THE ROLE OF THE VAGINAL MICROBIOME IN HPV INFECTION AND CARCINOGENESIS

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Alessandra Christina Gill

February 2018

Dedication

To my brother Roman who always believed in me.

Life has to end. Love doesn’t.

ABSTRACT

THE ROLE OF THE VAGINAL MICROBIOME IN HPV INFECTION AND CARCINOGENESIS by Alessandra Christina Gill

Cervical cancer is one of the most common cancers affecting women worldwide and is caused by persistent infection with high-risk types of human papillomavirus (HR-HPV). Cervical cancer rates are particularly high in developing regions, including South Africa, which currently also has the highest HIV prevalence anywhere in the world. The immunosuppression associated with HIV infection results in an increased likelihood of persistent HR-HPV infection in HIV-positive women, resulting in an increased risk of progression to cervical cancer, a risk that increases as HIV infection progresses.

Another risk factor for HR-HPV infection is a condition called bacterial vaginosis (BV) which is typified by a vaginal microbiome (VMB) made up of a diverse array of anaerobic bacteria with low numbers of lactobacilli. However, little detail is known about the bacterial species that are involved and the mechanisms that underlie this association. Using samples collected by the HARP (HPV in Africa Research Partnership) study, which was coordinated by the London School of Hygiene and Tropical Medicine, this study aimed to determine the association between the type of VMB and high-risk HPV infection and the presence of precancerous lesions of the cervix in HIV-infected South African women. This was achieved by characterising the VMB with the help of 16S rRNA gene sequencing of the V3-V4 region on the Illumina HiSeq platform, allowing identification of the bacterial taxa present in each vaginal sample. Laboratory and computational methods were optimised prior to sequencing of clinical samples to optimise the information gained.

We were able to determine that samples collected during the HARP study in the fixative medium BoonFix® and stored at room temperature were suitable for microbiome analysis. Furthermore, when using the Qiagen Blood and Tissue Kit, the inclusion of the collected vaginal swab in the proteinase K digestion step significantly increased DNA yield, which was correlated with lower levels of contaminant reads in the sequencing results. In order to further refine the sequencing results, we tested two newer OTU clustering methods (DADA2 and Swarm) against USEARCH and were able to show that newer methods allow better

species separation than those (such as USEARCH) that rely on a similarity threshold.

We found that this population of HIV-positive South African women had VMB profiles typical for women with black ethnic background and HIV infection: the majority had a *Lactobacillus iners*-dominated community, had a microbiome consisting of a diverse mixture of anaerobes typically associated with BV, or lay somewhere between these two extremes having lower levels of *L. iners* together with BV-associated anaerobes. In contrast, community types dominated by *L. crispatus*, *L. jensenii*, *Gardnerella vaginalis*, *Atopobium vaginae* or *Bifidobacterium* spp., or containing a relatively high abundance of pathobionts such as streptococci, staphylococci or *Enterobacteriaceae*, were uncommon or rare. Our results provide evidence for an association between HR-HPV infection and a high diversity vaginal microbiota typical of BV with a paucity of lactobacilli in general but especially *L. crispatus*. However, the effect sizes were relatively small in comparison to other studies, which may be due to the fact that our entire study population was HIV-positive: these women had a high baseline prevalence of non-lactobacilli-dominated VMBs (43% had a Nugent score of 7-10) and were more vulnerable to acquisition and persistence of pathogens (due to a median CD4 count of only 428 cells/µl) than HIV-negative women in the general population.

In multivariable models, we found no evidence of an association between the VMB and histological precancerous changes of the cervix in this cohort, beyond that related to persistent HR-HPV infection. The results of this study suggest that a high diversity vaginal microbiome with a paucity of lactobacilli is associated with HR-HPV infection in HIV-positive women and highlight the importance of taking HIV status into account when researching the VMB in HR-HPV infection and associated precancerous changes. A better understanding of how the vaginal microbiome impacts on the natural history of HPV infection and cervical cancer in women living with HIV could ultimately lead to improved management and treatment of these conditions in this high-risk group.

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Professor J. van de Wijgert without whom this project would not have been possible and Dr A. Darby who kept his sense of humour when I had lost mine. I must also thank the HPV in Africa Research Partnership coordinated by Professor P. Mayaud who kindly provided the clinical samples and metadata for my project, in particular Dr A. Chikandiwa for helping me organise the sample shipments. I am indebted to my supervisors, as well as Professor P. Mayaud, Professor S. Delany-Moretlwe and Dr S. Francis for their input into the study plan and both my supervisors for their guidance and careful consideration of my thesis drafts. I gratefully acknowledge the funding provided by the Institute of Infection and Global Health, and the provision of a maternity stipend.

I am grateful to have been adopted by the Darby lab group who have been a constant source of moral support and technical help, especially Fran Blow who taught me the ropes and Sam Whiteford for his help with bioinformatics software. I must also thank all the people who have shared a lab with me at various time points without ever complaining about my boxes being everywhere. In particular, Marijn Verwijs, Laura Goodfellow and Christina Bronowski - I have enjoyed our discussions. I am also thankful for the assistance of our technical staff who helped me with various aspects of the project, in particular Debby Sales and Caroline Broughton who never complained about the high volume of questions directed at them. I must also thank the staff from the Centre for Genomics Research for sharing their equipment and expertise, especially Luca Lenzi for his guidance with Swarm. I also thank my MSc buddy Helen Swift for some last-minute proof reading.

I was very lucky to have had the advice of several mentors during my PhD. In particular, the advice of Professor T. Skerry, Dr B. Makepeace and Professor M. Iturriza-Gomara has been invaluable. I must also thank my advisors, Professor N. French and Dr J Fothergill for their advice and comments on PhD progression.

Thanks also go to my family for their love and support during all the highs and lows, particularly my husband Richard, who had to pick up the pieces more than once and my parents who always encouraged me to achieve my potential. Finally I must mention my little Kaitlyn – my PhD baby – who has patiently shared me with my studies and my baby-to-be who gave me the motivation to finish my degree, fast.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PUBLICATIONS

Gill C, Van de Wijgert JH, Blow F, Darby AC (2016). Evaluation of Lysis Methods for the Extraction of Bacterial DNA for Analysis of the Vaginal Microbiota. PloS One 11:e0163148.

ABBREVIATIONS

AGC Atypical glandular cells

AIDS Acquired immunodeficiency syndrome

AIS Endocervical adenocarcinoma in situ

ART Antiretroviral therapy

ASC-H Atypical squamous cells

ASC-US Atypical squamous cells of undetermined significance

BV Bacterial Vaginosis

BVAB Bacterial vaginosis-associated bacterium

CIN Cervical intraepithelial neoplasia

DGGE Denaturing gradient gel electrophoresis

HARP HPV in Africa Research Partnership

HIV Human Immunodeficiency Virus

HPV Human Papillomavirus

HR-HPV High-Risk Human Papillomavirus

HSIL High grade squamous intraepithelial lesion

IARC International Agency for Research on Cancer

LR-HPV Low-Risk Human Papillomavirus

LSIL Low grade squamous intraepithelial lesion

OR Odds ratio

OTU Operational taxonomic unit

NMDS Nonmetric multidimensional scaling

LEfSe Linear discriminant analysis effect size

PCR Polymerase chain reaction

PERMANOVA Permutational multivariate analysis of variance

PVL Plasma viral load

STI Sexually transmitted infection

T-RFLP Terminal restriction fragment length polymorphism

UPGMA Unweighted Pair Group Method with Arithmetic Mean

VIA Visual inspection with acetic acid

VILI Visual inspection with Lugol's iodine

VMB Vaginal Microbiome

VMB-HARP vaginal microbiome substudy of the HPV in Africa Research Partnership study

CHAPTER 1: General Introduction

1.1 What is a Microbiome?

Although the term "microbiome" is often attributed to Nobel laureate Joshua Lederberg, the term (in its current context) was already in use much earlier (Whipps et al 1988). Lederberg defined the human microbiome as "the microbes that share our body space and that inhabit our skin, our mucous membranes, and our gut" (Lederberg 2001), though earlier definitions also included the environment inhabited by these microbes. Today, the precise definition of the word is a topic of debate (Schaechter 2013). Some researchers have used "microbiome" to describe the sum total of genomes associated with that microbiota (Ma et al 2012), and may include in that definition the surrounding environment (Cho and Blaser 2012). In that case, a "microbiome" can be thought of as the microbiological equivalent of a "biome", an ecological concept that describes a major type of ecosystem. However, in this text, I have elected to use the definition given by Lederberg, which can be viewed as an amalgamation of the word "microbe" with the suffix -ome (Lederberg and McCray 2001) and as such can be considered synonymous with "microbiota".

Although the human microbiome comprises viruses and eukaryotes as well as bacteria, the latter have been studied most extensively. Bacteria make up the largest portion of the microbiome in terms of biomass and there are well established methods available to study them. The majority of these bacteria reside in the large intestine where a single gram of dry faecal matter contains approximately 10^{11} bacterial cells (Franks et al 1998). This is an almost inconceivably large number, whose magnitude can be better appreciated when considering that this single gram of faeces contains more bacterial cells than there are human beings on earth. In terms of genetic content, it has been estimated that the intestinal microbiome contains roughly 2-4 million different genes, outnumbering their human host's genes by a factor of 50-100 (Hooper and Gordon 2001). Thus the metabolically active microbiome provides the host with access to gene products that are not within the repertoire of the human genome - thereby having huge potential to affect the health of the host. Although the intestinal microbiome is the largest population of microorganisms associated with the human body (Feeney and Sleator 2012), other regions of the body harbour sizable bacterial populations of their own. This includes other parts of the gastrointestinal tract, upper respiratory tract, skin, hair and vagina. The microbiome composition is different at each of these anatomical sites and, within the same individual, variation is greater between sites than temporal variation

at the same site (Cho and Blaser 2012). Microbiome composition is affected by the local environment, which is in turn affected by the life-stage and life-style of the human host. This study will focus on the vaginal microbiome (VMB), for which certain bacterial communities have been implicated as risk factors for the acquisition of sexually transmitted infections, the onward transmission of HIV (Cone 2014) and other reproductive health outcomes (van de Wijgert et al 2014). Hence, a better understanding of the VMB has great potential for the development of interventions that could lead to improvements in women's health.

According to Lederberg and McCray (2001), the term "microbiome" invites the consideration of the microbiota occupying a particular niche as a single entity, rather than separate units that function independently of one another. Until relatively recently, accurately describing the microbiome as a whole was impossible to achieve, due to the fact that a large proportion of microbes cannot be grown in the laboratory, making them impossible to identify by conventional microbiology techniques. As a result, organisms that are fastidious have perhaps not received the attention they deserve, while others that are easily cultured have dominated microbiological research. Good examples of this are *Escherichia coli*, the archetypical intestinal bacterium which, despite its prominence in the literature, actually only represents a small fraction of the gut microbiome (Arumugam et al 2011) and the discovery that *Lactobacillus iners* is the major constituent of the VMB in many women, but was missed in earlier studies as it fails to grown on standard culture media (Burton et al 2003). The advent of DNA sequencing in the 1970s opened the door for the development of new methods to identify microbes based on their genetic material, without the need for *in vitro* cultivation. However, early sequencing methods were costly and labour-intensive and lacked sensitivity, making them unsuitable for comprehensive analyses of the microbiome. Ambitious sequencing projects such as the Human Genome Project have driven the development of more efficient technologies which, shortly after the turn of the century, gave rise to modern high-throughput sequencing technologies that are capable of reading several billion bases in no more than a few days. As is described in detail later, the work that follows was completed using the next generation Illumina MiSeq sequencing platform to characterise the composition of the VMB and its association with human papillomavirus infection and cervical cancer in women infected with HIV.

1.2 The Vaginal Microbiome

1.2.1 Early work on the vaginal microbiome

Interest in the potentially beneficial nature of the normal vaginal microbiota was kindled by the pioneering work of Albert Döderlein in the late 19th and early 20th centuries (Döderlein 1897). He examined the vaginal microbiota of postpartum women microscopically and found that the microbiota in most healthy women contained significant numbers of a 'vaginal bacillus' (later named Döderlein's bacillus and now known as *Lactobacillus*). By contrast, other types of bacteria dominated the microbiota of women who developed puerperal fever. Döderlein was the first to suggest a link between the production of lactic acid by lactobacilli and protection from infection. He also warned against prophylactic disinfection of the vagina in the peripartum period, suggesting that it may lead to disruption of the normal flora which could allow pathogens to flourish (Schultz 1961).

In the decades that followed this work, studies to determine the association between the VMB and health and disease were hampered by the fact that the majority of bacteria could not be cultured, making it difficult to accurately describe the microbial population as a whole. Furthermore, methods to define what constitutes a normal or abnormal microflora were highly subjective leading to conflicting results (Amsel et al 1983, Martin 2012). Nonetheless, the existence of a clinical syndrome characterised by a thin homogenous vaginal discharge and associated with a loss of lactobacilli was soon recognised. This has been given a variety of names, initially being called "white discharge syndrome", later "nonspecific vaginitis" and finally "bacterial vaginitis", which was later changed to "bacterial vaginosis" in recognition of the fact that most cases are not accompanied by significant inflammation (Martin 2012).

Standardised criteria for the diagnosis of bacterial vaginosis (BV) were first published in the 1980s by Amsel and co-workers (Amsel et al 1983), who aimed to create a method that could easily be used by medical practitioners. The "Amsel criteria" for the diagnosis of BV are satisfied by the presence of three out of the following four conditions: (1) a vaginal pH above 4.5, (2) the presence of a thin homogenous vaginal discharge, (3) the production of a fishy odour on application of potassium hydroxide (which is indicative of the presence of a significant amount of amines, mainly putrescine and cadaverine, produced by certain bacteria) and (4) the presence of clue cells (vaginal epithelial cells covered with adherent bacteria) under the microscope (Amsel et al 1983). The Amsel criteria provide an indirect

measure of the vaginal microbiota and are simple to assess in a clinical setting. However, although they have contributed to the study of the vaginal microbiota, they suffer from significant limitations. First, recent studies have shown that, with the exception of pH, they do not correspond well with bacterial profiles as determined by molecular methods (van de Wijgert et al 2014). Second, all but the measurement of vaginal pH are highly subjective causing concern that they could introduce a high degree of observer bias to clinical research studies, particularly for larger multi-centre projects. There was therefore a need for a standardised laboratory-based test.

In 1991, Nugent and co-workers published revised criteria (now known as the Nugent score) for the evaluation of vaginal smears by Gram staining (Nugent et al 1991). This built on earlier work by Spiegel and others, who had developed a binary test for BV that was based on the assessment of gram-stained vaginal smears, and which was developed by using the Amsel criteria as a gold standard (Spiegel et al 1983). Rather than being a binary test, the Nugent score ranges from 0-10 (see Table 1.1), with a lower score for the presence of large gram-positive rods (*Lactobacillus* morphotype), and a higher score for the presence of small gram-variable rods, gram-negative rods and curved gram-variable rods (representing *Gardnerella*, *Bacteriodes* and *Mobiluncus* morphotypes, respectively) (Nugent et al 1991). These morphotypes were chosen due to their high intra- and inter-centre repeatability, and the resulting score was weighted according to expert opinion at the time (Nugent et al 1991). As expected, the Nugent score has been shown to have high inter-observer reliability (Forsum et al 2002, Zarakolu et al 2004). It has been extensively used for the diagnosis of BV in clinical research studies, which has allowed the identification of a clear relationship between the VMB and various adverse health outcomes including pelvic inflammatory disease (Haggerty et al 2004), preterm birth (Hillier et al 1995) and infection with sexually transmitted diseases such as HIV (Martin et al 1999).

Although BV is considered a clinical condition in its own right that may present with a malodorous vaginal discharge and pruritus, it should be noted that the majority of women with BV diagnosed by either Nugent score or Amsel criteria do not report any vaginal symptoms at all. Furthermore, vaginal symptoms are only marginally less frequently reported by women classified as having normal microflora by these methods (Amsel et al 1983, Klebanoff et al 2004, Koumans et al 2007). This may at

least in part be due to a lack of consensus amongst both patients and clinicians as to what constitutes an ordinary amount, consistency and odour of vaginal secretions (Anderson et al 2004). When assessed by a healthcare practitioner, malodour is fairly specific for BV by Nugent score, but a large proportion of women with BV are not classified as having any malodour, even after the addition of potassium hydroxide to vaginal secretions (Beverly et al 2005, Simoes et al 2006). In contrast, when using the Nugent score as the gold standard, the sensitivity and specificity of a 'characteristic' vaginal discharge are greatly variable between studies (Beverly et al 2005, Simoes et al 2006), which may reflect the highly subjective nature of this assessment.

Table 1.1 Calculation of the Nugent score. Gram-stained vaginal smears are examined under oil immersion and three bacterial morphotypes are quantified per high-power field: *Lactobacillus* morphotypes (large gram-positive rods), *Gardnerella* and *Bacteriodes* morphotypes (small gram-variable rods and gram-negative rods respectively) and *Mobiluncus* morphotypes (curved gram-variable rods). Numbers of each morphotype are converted to a numerical score (see left-hand column) and added to give the final Nugent score. A score of 0-3 is considered normal, a score of 4-6 is considered intermediate and a score of 7-10 is considered consistent with bacterial vaginosis. Adapted from Nugent et al (1991).

Score	*Lactobacillus* morphotype	*Gardnerella/Bacteriodes* morphotype	*Mobiluncus* morphotype
0	>30	none	none
1	5-30	<1	≤4
2	1-4	1-4	≥5
3	<1	5-30	-
4	none	>30	-

1.2.2 Modern methods to study the vaginal microbiome

The advent of DNA sequencing technologies has revolutionised the study of microbial populations as it makes it possible to identify bacteria without the need to first culture them in the laboratory. These "molecular methods" were first used to study the VMB at the beginning of this century (Burton and Reid 2002). Although a number of different methods have been used they are all based on the same principle and involve extraction of genomic DNA from samples, usually followed by PCR-based amplification and sequencing of a universal bacterial gene, i.e. a gene that is invariably present in all bacterial species. The most commonly used gene is that encoding 16S ribosomal RNA, which is particularly suited to this work since it is not only universal, but has evolved at a relatively slow rate. This means that it contains regions that are highly conserved between distantly related organisms, allowing binding of universal primers and comprehensive PCR amplification even from populations with high species diversity. Bacterial taxa can then be identified

based on the more variable intervening regions of the gene known as hypervariable regions (Case et al 2007).

The first molecular techniques involved the use of Sanger sequencing, which relies on being able to generate a large amount of amplicon from a single species. To achieve this, amplicons were initially separated by denaturing gradient gel electrophoresis (DGGE), allowing identification of species that were numerically abundant enough to produce bands on a gel (Burton and Reid 2002). Better resolution could be achieved by amplifying genes in a bacterial vector prior to sequencing (Burton et al 2004). Other approaches such as terminal restriction fragment length polymorphism analysis (T-RFLP) or microarray analysis have also been used instead of sequencing, but researchers are then limited to known bacterial species (Borgdorff et al 2014, Coolen et al 2005). The use of Sanger sequencing provided the first opportunity to identify unculturable vaginal bacteria, but still lacked the resolution to identify minority species.

Over the last decade, the increasing affordability of high-throughput sequencing technologies has made it possible to produce a vast amount of sequencing data to describe the VMB. Initial studies utilised 454 FLX pyrosequencing technology (Spear et al 2008), as other methods were not able to produce long enough DNA sequences, or "reads". However, continued technological improvements have now made it possible to accurately describe the VMB at a relatively lower cost with the use of Illumina technology (Fadrosh et al 2014), which has been used in this study. These next-generation sequencing technologies are for the first time allowing detailed description of vaginal communities, with great potential to improve our understanding of the health implications of distinct VMB structures.

1.2.3 The composition of the vaginal microbiome

Despite the current view that a vaginal bacterial community dominated by *Lactobacillus* spp is optimal, recent work has shown that a substantial proportion of apparently healthy women possess a VMB made up of a diverse community of other facultative anaerobes and strict anaerobes. In some ethnic groups, this diverse VMB-type can have a prevalence as high as 40% (Ravel et al 2011). It follows that although in some cases a diverse VMB can be described as "normal", it may not provide optimal protection from negative health outcomes.

There are now a number of studies that have used culture-independent techniques to characterise the VMB in healthy women of reproductive age and those with bacterial vaginosis. Most of these have identified groups of women whose VMB is dominated by a single taxon. Often this taxon is *Lactobacillus crispatus* or *L. iners*, but VMB communities dominated by other taxa, including *Lactobacillus gasseri*, *Lactobacillus jensenii* and *Gardnerella vaginalis* have also been identified (van de Wijgert et al 2014). Furthermore, VMB types that are not dominated by a single taxon are also prevalent. These communities are variable and consist of a diverse assortment of facultative and strict anaerobes that commonly include *Atopobium*, *Dialister*, *Gardnerella*, *Megasphaera*, *Prevotella* and *Sneathia* species. A more extensive list of taxa that have been identified in vaginal samples is presented in Table 1.2.

Unsurprisingly, women with BV by Nugent score have a distinct VMB profile compared to women without BV. The vast majority of women without BV have VMB communities dominated by *Lactobacillus* species (Srinivasan et al 2012). Interestingly, while *L. crispatus* is mainly found in women without BV, *L. iners* is also common in subjects with BV, but may be more abundant in women without the condition (Srinivasan et al 2012). In contrast, women with BV by Nugent score tend to have VMB communities with higher overall bacterial diversity (van de Wijgert et al 2014), which in turn correlates positively with an increase in vaginal pH (Drell et al 2013, Human Microbiome Project Consortium 2012). Several bacterial species have been associated with a high Nugent score, including *G. vaginalis*, *Atopobium vaginae*, *Leptotrichia*/*Sneathia* spp, *Megasphaera* spp. and *Prevotella* spp. Often not only the presence, but also an increased abundance of these species is predictive of BV (Datcu et al 2013, Ling et al 2010, Ling et al 2013, Yeoman et al 2013). Considering that the Nugent score is based on the semi-quantitative assessment of selected bacterial morphotypes (see Table 1.1), it should be no surprise that there are strong associations between the Nugent score and bacterial community composition profiles as determined by molecular techniques. Interestingly, two studies that defined BV based on the Amsel score, did not identify a significant association of *Lactobacillus* spp. with BV (Mitchell et al 2009, Yeoman et al 2013), while two other studies did find an association (Haggerty et al 2009, Shipitsyna et al 2013). These differences may be due to the small sample sizes or the subjective nature of the Amsel score, but it is interesting to note that one of the

Table 1.2 Taxonomy and selected characteristics of bacteria commonly reported in vaginal samples (Aagaard et al 2012, Chaban et al 2014, Datcu et al 2013, Dols et al 2011, Drell et al 2013, El Aila et al 2009, Fettweis et al 2014, Forney et al 2010, Frank et al 2012, Gajer et al 2012, Hernández-Rodríguez et al 2011, Hickey et al 2013, Huang et al 2015, Hummelen et al 2010, Kim et al 2009, Lee et al 2013, Ling et al 2010, Martin et al 2012, Oakley et al 2008, Pépin et al 2011, Ravel et al 2011, Shipitsyna et al 2013, Smith et al 2012, Spear et al 2008, Srinivasan et al 2012, Wertz et al 2009, Yamamoto et al 2009, Zhou et al 2010). Bacterial classification for *Mageeibacillus indolicus* (formerly "BVAB3") as described by Austin et al (2015) and for the remaining BVAB species as described by Fredricks et al (2005). Remaining information is taken from Bergey's Manual of Systematic Bacteriology (Brenner et al 2005, DeVos et al 2009, Krieg et al 2010).

BACTERIA WITH GRAM POSITIVE PHYLOGENY COMMONLY REPORTED AS PART OF THE VAGINAL MICROBIOME (continued on next page)

Phylum	Class	Order	Family	Genus/species	Gram stain[1]	Anaerobe/aerobe	Morphology
Firmicutes (cont.)	Negativicutes	Selenomonadales	*Veillonellaceae*	*Dialister*	-	obligate anaerobe or microaerophile	coccobacillus
				Megasphaera	-	obligate anaerobe	coccus
				Veillonella	-	obligate anaerobe	coccus
Tenericutes	Mollicutes	Mycoplasmatales	*Mycoplasmataceae*	*Mycoplasma*	-	facultative anaerobe	pleomorphic
				Ureaplasma	-	facultative anaerobe	coccus/coccobacillus

BACTERIA WITH GRAM NEGATIVE PHYLOGENY COMMONLY REPORTED AS PART OF THE VAGINAL MICROBIOME

Phylum	Class	Order	Family	Genus/species	Gram stain[1]	Anaerobe/aerobe	Morphology
Bacteroidetes	Bacteroidia	Bacteroidales	*Bacteroidaceae*	*Bacteroides*	-	obligate anaerobe	rod
			Porphyromonadaceae	*Porphyromonas*	-	obligate anaerobe	coccobacillus
			Prevotellaceae	*Prevotella*	-	obligate anaerobe	short rod
Fusobacteria	Fusobacteriia	Fusobacteriales	*Leptotrichiaceae*	*Leptotrichia*	-	anaerobe	rod
				Sneathia	-	anaerobe, moderately aerotolerant	long rod
Proteobacteria	Gammaproteobacteria	Enterobacteriales	*Enterobacteriaceae*	*Escherichia*	-	facultative anaerobe	rod

[1]Gram staining characteristics given as negative (-), positive (+), variable (+/-) or unknown (?).

BACTERIA WITH GRAM POSITIVE PHYLOGENY COMMONLY REPORTED AS PART OF THE VAGINAL MICROBIOME (continued from previous page)

Phylum	Class	Order	Family	Genus/species	Gram stain[1]	Anaerobe/aerobe	Morphology
Actinobacteria	Actinobacteria	Actinomycetales	*Actinomycetaceae*	*Mobiluncus*	- or +/-	obligate anaerobe	curved rod
		Bifidobacteriales	*Bifidobacteriaceae*	*Bifidobacterium*	+	anaerobe (some facultative or aerotolerant)	rod
				Gardnerella vaginalis	- or +/-	facultative anaerobe (some strains obligate)	small pleomorphic rod
		Coriobacteriales	*Coriobacteriaceae*	*Atopobium vaginae*	+	facultative anaerobe	coccobacillus
				Eggerthella	+	obligate anaerobe	rod
Firmicutes	Bacilli	Bacillales	*Staphylococcaceae*	*Staphylococcus*	+	facultative anaerobe	coccus
			unassigned	*Gemella*	+	facultative anaerobe	coccobacillus
		Lactobacillales	*Aerococcaceae*	*Aerococcus*	+	facultative anaerobe	coccus
			Lactobacillaceae	*Lactobacillus*	+	facultative (or obligate) anaerobe	large rod
			Streptococcaceae	*Streptococcus*	+	facultative anaerobe	coccus
	Clostridia	Clostridiales	?*Lachnospiraceae*	"BVAB1"[2]	?	anaerobe	curved rod
				"BVAB2"[3]	?	anaerobe	short rod
			Peptostreptococcaceae	*Peptostreptococcus*	+	obligate anaerobe	coccus
			Ruminococcaceae	*Mageeibacillus indolicus*	-	obligate anaerobe	rod
			unassigned	*Anaerococcus*	+	obligate anaerobe	coccus
				Finegoldia	+	obligate anaerobe	coccus
				Parvimonas	+	obligate anaerobe	coccus
				Peptoniphilus	+	obligate anaerobe	coccus

[1]Gram staining characteristics given as negative (-), positive (+), variable (+/-) or unknown (?).
[2]Bacterial vaginosis associated bacterium 1
[3]Bacterial vaginosis associated bacterium 2

studies also found that the metabolomic profiles were much more closely related to the Amsel criteria than to the Nugent score, indicating that the Amsel score may better reflect community function (Yeoman et al 2013). The wealth of information offered by next-generation sequencing studies may in future refine our understanding of the VMB community profile and function that most accurately reflects symptomatic BV.

Data on the stability of VMB composition within the same subject over time is still relatively sparse. However, a comprehensive longitudinal study of 32 healthy women of reproductive age was carried out by Gajer et al (2012). Samples were collected at weekly intervals over a period of 16 weeks and were analysed using pyrosequencing. The results showed that some women have relatively stable community profiles, which was particularly true for communities dominated by *L. crispatus* and *L. gasseri*, but the study also identified four individuals with a high diversity VMB and persistent asymptomatic BV by Nugent score that also had very stable community profiles. In women with stable VMB profiles, the most significant fluctuations in community structure usually coincided with menses, with sexual activity having a lesser negative impact on stability. In contrast, other communities showed marked shifts in composition over short periods of time and the authors hypothesised that these communities might exist in a limited number of alternative equilibrium states. Shifts in community structure were dependent on the type of vaginal microbiome. Communities dominated by *L. crispatus* most often switched to either an *L. iners*-dominated community or a community with moderate numbers of lactobacilli in combination with various species of strict anaerobes. Conversely, *L. iners*-dominated communities also switched to the latter, but were twice as likely to switch to a high diversity state typical of BV. Additionally, most conversions from this high diversity state usually resulted in an *L. iners*-dominated VMB. Another host factor that is associated with a particular VMB composition is pregnancy, in which the VMB is generally more stable and of lower diversity and species richness - with higher levels of Lactobacillus spp and lower levels of BV-associated bacteria - when compared with non-pregnant women (Aagaard et al 2012, Jespers et al 2015, Romero et al 2014, Walther-António et al 2014).

1.3 Human Papillomavirus - a Public Health Problem

1.3.1 HPV and its effects on human health

Human papillomaviruses (HPVs) are small double-stranded DNA viruses that lack an envelope. More than 200 different genotypes are formally distinguished and are

separated on the basis of having less than 90% similarity in DNA sequence of the L1 gene (the gene encoding the major capsid protein) when compared with any other HPV type. Viruses that differ by less than that are referred to as subtypes or variants (Burk et al 2013, Bzhalava et al 2015).

HPVs infect and replicate in keratinocytes, the epithelial cells of the skin (for cutaneotropic HPV types) and mucosal sites such as the cervix (for mucosotropic HPV types). The virions require access to the basal layer of epithelial cells in order to establish infection, probably gaining entry through microabrasions in the skin or mucosal surface (Schiller et al 2010). HPVs are currently grouped into five major genera, such that members of each genus differ in their L1 open reading frame by more than 60% from members of other genera. These are further divided into species which share 60-70% sequence similarity (Bzhalava et al 2015). The sexually-transmitted genital HPV types are mucosotropic and belong to the alphapapillomavirus genus, of which they make up the majority. Genital HPVs are further divided into low-risk and high-risk types according to their carcinogenic potential. In the general population, infection with low-risk HPV (LR-HPV) types (which also include the cutaneotropic alphapapillomaviruses) is usually either asymptomatic or characterised by the appearance of benign papillomas. They are only rarely associated with cancer (Doorbar et al 2012).

By contrast, high-risk HPV (HR-HPV) can be identified in virtually all carcinomas of the cervix, and they are therefore considered a necessary instigating factor for the development of cervical cancer (Doorbar et al 2012, Walboomers et al 1999). The most commonly encountered histological type of cervical cancer is squamous cell carcinoma, which arises from the squamous epithelium of the ectocervix and represents 75-90% of cases. The largest proportion of remaining cases are adenocarcinomas, which arise from the glandular epithelium of the endocervix (Mathew and George 2009, Tjalma et al 2005). HR-HPV infection has also been associated, albeit less strongly, with other anogenital cancers (Wakeham and Kavanagh 2014) and a number of head and neck tumours (Leemans et al 2011). However, most HR-HPV infections are eventually cleared within 6-12 months of infection and do not result in cancer (Zur Hausen 2002). In fact, infection with HPV is extremely common worldwide, with an average prevalence of 11.7% in women with normal vaginal cytology. There is regional variation in infection prevalence, with sub-Saharan Africa having the highest HPV prevalence at 24.0%. Globally, the most

commonly encountered HPVs are the high-risk types 16 and 18, (Bruni et al 2010), which also have the strongest association with cervical cancer (IARC 2012).

Table 1.3 Human alphapapillomaviruses types currently recognised by the HPV Reference Centre (see http://www.hpvcenter.se/html/refclones) and their corresponding IARC category: group 1 (carcinogenic to humans), group 2A (probably carcinogenic to humans) and group 2B (possibly carcinogenic to humans). The remainder are designated as group 3 - not classifiable (IARC 2012).

Alpha-1	HPV32	
	HPV42	
Alpha-2	HPV3	
	HPV10	
	HPV28	
	HPV29	
	HPV77	
	HPV78	
	HPV94	
	HPV117	
	HPV125	
	HPV160	
Alpha-3	HPV61	
	HPV62	
	HPV72	
	HPV81	
	HPV83	
	HPV84	
	HPV86	
	HPV87	
	HPV89	
	HPV102	
	HPV114	
Alpha-4	HPV2	
	HPV27	
	HPV57	
Alpha-5	HPV26	Group 2B
	HPV51	Group 1
	HPV69	Group 2B
	HPV82	Group 2B
Alpha-6	HPV30	Group 2B
	HPV53	Group 2B
	HPV56	Group 1
	HPV66	Group 2B
Alpha-7	HPV18	Group 1
	HPV39	Group 1
	HPV45	Group 1
	HPV59	Group 1
	HPV68	Group 2A
	HPV70	Group 2B
	HPV85	Group 2B
	HPV97	Group 2B
Alpha-8	HPV7	
	HPV40	
	HPV43	
	HPV91	
Alpha-9	HPV16	Group 1
	HPV31	Group 1
	HPV33	Group 1
	HPV35	Group 1
	HPV52	Group 1
	HPV58	Group 1
	HPV67	Group 2B
Alpha-10	HPV6	
	HPV11	
	HPV13	
	HPV44	
	HPV74	
Alpha-11	HPV34	Group 2B
	HPV73	Group 2B
	HPV177	
Alpha-13	HPV54	
Alpha-14	HPV71	
	HPV90	
	HPV106	

HPV 16 has by far the strongest association with cervical cancer, with one meta-analysis finding this type in 54.4% of cases of invasive cervical cancer, compared with only 2.6% of women with normal cervical cytology. There is also strong

evidence that HPV 18 is associated with cervical cancer, with 15.9% of cases of invasive cervical cancer being positive for this type, compared with 0.9% of controls (IARC 2012). Both HPV 16 and 18 are associated with squamous cell carcinoma and adenocarcinoma of the cervix, the predominant types of cancer at this site. However, HPV 18 is relatively more important as a cause of adenocarcinoma, causing similar numbers of cases as HPV 16 (Clifford et al 2003). In addition, there is convincing epidemiological and experimental evidence that other HPV types also cause cervical cancer, although they are less carcinogenic than HPV 16 and 18. Currently 12 human alphapapillomaviruses are classified as carcinogenic by the International Agency for Research on Cancer (IARC): 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58 and 59 (IARC 2012). Of these 16, 18, 31, 33, 35, 45, 52 and 58 are most often identified in cases of cervical cancer (Smith et al 2007). A further 13 HPV types are classed as probably or possibly carcinogenic by the IARC (see Table 1.3), based on weak evidence from epidemiological studies, limited *in vitro* experimental evidence, or a close phylogenetic relationship with known carcinogenic types (IARC 2012).

1.3.2 How does HPV cause cervical cancer?

As a causative agent of cervical cancer, HPV has the strongest association with cancer ever identified for any cancer-causing agent (Scheurer et al 2005). It is therefore not surprising that cervical cancer has been the best studied of the HPV-related cancers. Significantly, cervical cancer is also one of the most common cancers in women worldwide with an estimated 528,000 new cases every year and over 260,000 deaths annually. The majority of this burden falls on developing regions, including Sub-Saharan Africa, where cervical cancer is the most common cause of cancer-related death in women (Ferlay et al 2013). In this region, the age-standardised death rate for cervical cancer is 22.5 per 100,000 women. In the Republic of South Africa this rate is typical for the region at 18.0, much higher than that observed in Western Europe and the United States, where rates are below 3.0 (Ferlay et al 2013). These differences can be partly explained by a lack of awareness, insufficient access to suitable preventative cervical cancer screening programmes leading to later presentation of cancer cases, and inadequate access to treatment in developing countries (Denny et al 2014).

The mechanism by which HPV causes malignant transformation relates in part to its molecular biology, a brief summary of which is provided below. After infection of a basal keratinocyte, the HPV life cycle is closely related to differentiation of the host

cell. Normally, epithelial cells undergo terminal differentiation as they progress towards the superficial layers of the epithelium, i.e. they no longer multiply. HPV is dependent on the host cell's replication machinery to produce new viral DNA and proteins and has therefore evolved mechanisms to maintain the cell in a dividing state. This is achieved through the induction of host cell expression of various viral proteins. The HPV genome is approximately 8000 base pairs in size and encodes 8 well-defined open reading frames: E1, E2, E4, E5, E6 and E7, which are first expressed at an early stage of infection and L1 and L2, which encode capsid proteins that are expressed later (Lehoux et al 2008). In HR-HPV types, proteins E6 and E7 in particular have been found to have an important role in the progression of infection to cancer. These proteins have been most extensively studied in HPV 16 and 18. In these viruses, a major function of E7 is to inactivate the tumour suppressor protein pRb, thereby maintaining the cell in a proliferative state (Lehoux et al 2008). E6 acts synergistically with E7 to cause cancer by ensuring cell survival. This is achieved by E6-mediated inactivation of the tumour-suppressing regulatory protein p53, thus preventing programmed cell death that would otherwise be triggered by the abnormal cellular growth caused by E7 (Hiller et al 2006). Additionally, E6 activates telomerase which extends the reproductive lifespan of the cell, favouring the development of cancer (Xu et al 2008).

During the normal HPV infection cycle, viral DNA is maintained within the nucleus separate from host genomic DNA in episomal form. However, in rare cases the viral genome becomes integrated into the host's chromosomal DNA. This may facilitate progression to cancer if it disrupts the E2 open reading frame, leading to the loss of suppression of viral oncoproteins by E2. In the majority of cervical cancers, HPV is found integrated into the host genome, but this is not an essential event for the development of cancer (Pett and Coleman 2007).

Consistent with the observation that only a small fraction of HR-HPV infections progress to cervical cancer, the overexpression of E6 and E7 alone is insufficient to cause malignant transformation. The mechanisms that ultimately result in cancer are complex and involve the accumulation of several genetic alterations. Although mutations are rare in normal cells, HPV infection encourages the process by causing genomic instability, at least in part through disruption of DNA repair processes and by preventing apoptosis of abnormal cells (Tommasino 2014). These changes usually occur over the course of several years, explaining why persistent

infection with a HR-HPV type, permitted by effective immune evasion, is a prerequisite for cervical carcinogenesis (Lehoux et al 2008).

Table 1.4 Summary of the most commonly reported cytology results according to the 2001 Bethesda system. Other findings, such as infection and the presence of endometrial cells, may also be included (Solomon et al 2002).

BETHESDA CYTOLOGY CLASSIFICATION	DESCRIPTION
Negative for intraepithelial lesion or malignancy	
Abnormal squamous epithelial cells	
Atypical squamous cells of undetermined significance (ASC-US)	Cells appear abnormal which may reflect HR-HPV infection, but may also have other causes. A small proportion of women with ASC-US have CIN grades 2 or 3.
Atypical squamous cells (ASC-H)	Cells appear abnormal and a high-grade squamous intraepithelial lesion cannot be ruled out. May be more likely to have CIN grades 2 or 3 compared to women with ASC-US.
Low grade squamous intraepithelial lesion (LSIL)	Low grade changes which are most likely to reflect transient HR-HPV infection; thought to reflect CIN grade 1
High grade squamous intraepithelial lesion (HSIL)	High grade changes which are most likely to reflect persistent HR-HPV infection; thought to reflect CIN grades 2 or 3
Squamous cell carcinoma	Highly abnormal cells that likely represent invasive cancer
Abnormal glandular (endocervical or endometrial) cells	
Atypical glandular cells (AGC)	Cells appear abnormal which may reflect HR-HPV infection. Tends to be more often associated with high grade lesions than ASC-US.
Atypical endocervical cells, favour neoplastic	Cells appear abnormal with some changes that are suggestive of a high grade lesion.
Endocervical adenocarcinoma in situ (AIS)	High grade changes which are most likely to reflect adenocarcinoma in situ, but a proportion of cases will have invasive disease.
Adenocarcinoma	Highly abnormal cells that likely represent invasive cancer.

1.3.3 Grading of cervical changes caused by HR-HPV infection

Persistent HR-HPV infection may lead to changes in the cervical epithelium which start in the lower epithelial layers and progress until they involve the superficial epithelium. Over time, these pre-cancerous lesions can progress to invasive cancer. Staging of these changes is most accurately achieved by examination of histological specimens, but is often carried out less invasively by cytological examination of exfoliated cervical cells. However, the correlation between cytology results and histology of cervical biopsies (which is considered the gold standard) is far from perfect (Lonky et al 1999). The most common system used for reporting cytology

results is called the Bethesda System, which was updated in 2001 (Solomon et al 2002). Broadly speaking, women are classified into those who are negative for intraepithelial lesion or malignancy and those who have abnormalities of squamous epithelial or glandular cells, with the main purpose being to assess squamous epithelial cells (see Table 1.4). However, a definitive diagnosis of cervical neoplasia requires a subsequent histopathological examination of a cervical biopsy. There are currently three grades of histopathological abnormalities of the squamous epithelium: cervical intraepithelial neoplasia (CIN) grades 1, 2 and 3, developed from the system proposed by Richart and Barron (1969). This grading system

Table 1.5 Histopathological grading of cervical intraepithelial neoplasia (CIN).

HISTOPATHOLOGY CLASSIFICATION	DESCRIPTION
CIN1	Nuclear abnormalities are minimal, mitotic figures are few and undifferentiated cells are confined to the lower epithelial layer.
CIN2	Nuclear changes are more marked, and mitotic figures are more common with cellular dysplasia confined to the lower half of the epithelium.
CIN3	Nuclear dysplasia and mitotic figures are seen throughout the epithelium.

involves an assessment of nuclear abnormalities consistent with neoplastic change, prevalence of mitotic figures (which are indicative of high cellular division rates) and the thickness of epithelium that is dysplastic (see Table 1.5). Other cellular changes, such as those secondary to inflammation can complicate the diagnosis of CIN and reduce inter-observer agreement. However, accuracy can be improved by performing a consensus review in which difficult cases are assessed by more than one pathologist (De Vet et al 1995). Not all CIN lesions progress to invasive cancer, some persist and most regress. However, the higher the grade of CIN, the more likely it is that women will develop invasive cervical cancer and that this change will occur within a shorter period of time. Without taking differing follow-up time into account, one review of studies on CIN progression (as assessed by cytology or histology) estimated that regression occurred in approximately 57, 43 and 32% of cases for women diagnosed with CIN1, 2 and 3, respectively. On the other hand, progression to invasive cancer occurred in 1, 5 and 12% of cases, respectively (Östör 1993). To avoid confusion, it should be noted that some pathologists also use CIN terminology to classify cytology results and may report this instead of, or in addition to SIL.

1.3.4 HPV, cervical cancer and the vaginal microbiome

As previously mentioned, the human vaginal microbiota is thought to play an important role in the prevention or acquisition of sexually transmitted diseases. Several studies have found a positive association between BV (as determined by Amsel or Nugent score) and both HPV infection and cervical cancer, and this has been confirmed by the results of two systematic reviews and meta-analyses (Gillet et al 2011, Gillet et al 2012). Most of the reviewed studies were cross-sectional, making it difficult to draw conclusions about the temporality of this association. However, the results of two large longitudinal studies suggest that a Nugent score of 7 or above is a risk factor for incident HPV infection (King et al 2011, Watts et al 2005).

There are a handful of published studies that have used molecular techniques to investigate the association between the VMB and HPV infection. One study using a *Lactobacillus*-specific micro-array found that prevalent HIV and HPV infection were less common in women with *L. crispatus*-type dominated flora and more common in *L. brevis*-type dominated flora in South African women (Dols et al 2012). A further cross-sectional study on women in China that characterised the VMB using DGGE found an association between higher bacterial diversity and HPV infection. Specifically, the detection of *L. gasseri* and *G. vaginalis* were associated with prevalent HPV infection, and there was a non-significant trend for an association with the presence of *A. vaginae*. In contrast the majority of VMB profiles dominated by *Lactobacillus gallinarum*-type (which could not be speciated more accurately, but probably represents *L. crispatus*) and *L. iners* were HPV negative (Gao et al 2013). A further cross-sectional study using a phylogenetic micro-array showed that in a group of 174 Rwandan female sex workers, those with *L. crispatus* dominated VMB (and to a lesser extent women with a *L. iners*-dominated VMB) were significantly less likely to have HPV infection than women with other types of VMB (Borgdorff et al 2014).

There are currently only five published studies that have utilised next generation sequencing to investigate the relationship between the VMB and HPV infection. The first of these was conducted by Lee and others, and was a cross-sectional study in a small cohort of Korean twins. The authors concluded that HPV positivity was associated with the presence of Fusobacteria (including *Sneathia* spp.), *Prevotella* spp. and Clostridiales. *Sneathia* spp. in particular were associated with high-risk

HPV infection. Sub-analysis of nine monozygotic twin pairs showed that the VMB of HPV-positive twins were more diverse with a significantly lower number of lactobacilli (particularly *L. iners*) and increased proportions of *Prevotella*, *Sneathia*, *Dialister* and *Bacillus* spp. Despite these differences, there was no significant relationship between the VMB community type (determined by clustering microbial profiles according to diversity and relative abundance of all detected bacterial phylotypes) and HPV infection status (Lee et al 2013). The second publication using next generation sequencing reported on a longitudinal study on a small cohort of US women. In this study, the VMB community type was significantly associated with HPV clearance at the next sampling time point 3-4 days later. However, there was no significant association with incident HPV infection. The community type dominated by *L. jensenii* was associated with the fastest clearance rate, while a high diversity lactobacillus-low VMB community type (including amongst others higher proportions of *Atopobium*, *Gardnerella*, and *Prevotella* spp.) was associated with the slowest clearance rate (Brotman et al 2014). A further study in 278 Nigerian women found differences in within-group beta diversity between HR-HPV positive and negative groups, but only among women who were HIV negative. The main differences between these groups were an increase in Leptotrichiaceae (which includes the genera *Sneathia* and *Leptotrichia*), Prevotellaceae, Clostridiaceae and Peptostreptococcaceae in the HR-HPV positive group (Dareng et al 2016). A further small study in HIV-positive and -negative women concluded that women with higher levels of *L. crispatus* have reduced HR-HPV infection levels, while women in the high diversity VMB tertile were significantly more likely to have HR-HPV (Reimers et al 2016). A very recent publication on a small group of Italian women found that a microbiome type dominated by variable proportions of *Gardnerella*, *Prevotella*, *Atopobium* and *Sneathia* was associated with persistence of HR-HPV infection at one-year follow-up. Of these genera, linear discriminant analysis effect size (LEfSe) analysis identified *Atopobium* as significantly increased in the women with HR-HPV persistence compared to women that cleared infection. Furthermore, women in the persistence group had significantly higher amounts of a *Gardnerella* sialidase gene involved in biofilm formation, compared to women who cleared infection. Additionally, *Sneathia*, *Megasphaera*, *Pseudomonas*, *Pediococcus* and *Brevibacterium* were enriched in women who were HR-HPV positive at baseline compared to those who were not (Di Paola et al 2017). Larger longitudinal studies are now needed to clarify the association of the VMB with incident HPV infection.

Four additional studies have employed molecular methods to investigate the relationship between the VMB and cervical cancer. One study on women in the UK found that the high diversity vaginal community type was proportionally more common with increasing vaginal cytological abnormalities, but this trend was not statistically significant (Mitra et al 2015). A similar study in Korean women found that having a predominance of *A. vaginae*, *L. iners* and *G. vaginalis* and low proportional abundance of *L. crispatus* (as determined by factor analysis) was associated with a higher risk for high grade cytology. When comparing the abundance of these species individually, only *A. vaginae* showed a significant difference in median abundance between groups. Women who had high grade cytology also had a higher *A. vaginae*:*L. iners* ratio, which could be confirmed by real-time PCR (Oh et al 2015). A further study in a small group of Mexican women found that higher cytological grades were associated with higher alpha-diversity and within-group beta diversity compared to HR-HPV negative women with normal cytology. Samples from women with normal cytology and no HR-HPV tended to contain predominantly *L. crispatus* and *L. iners*, whereas women with abnormal cytology tended to have higher proportions of *Sneathia* and *Fusobacterium* spp. (Audirac-Chalifour et al 2016). However, in light of the small sample number, these results should be interpreted with caution. Conversely, a larger study involving 430 US women with abnormal cytology results found that women with CIN2 and above were more likely to have a vaginal community dominated by *L. iners* and *L. crispatus* than women with CIN1 (considered normal). Similarly, LEfSe analysis found that women with CIN2 or higher had significantly higher levels of Lactobacillaceae, *Lactobacillus* spp. and *Lactobacillus reuteri* (Piyathilake et al 2016). It is possible the differences seen in this study could be partly due to the choice of control group. Further studies, particularly those of a longitudinal nature are needed to clarify the relationship of the VMB with CIN and to differentiate it from that with HR-HPV.

Currently, very little is known about the underlying mechanisms by which the composition of the VMB is linked to HPV infection and the progression to cervical cancer. However, potential mechanisms have been proposed. As previously described, a key step to infection with HPV is access of the virus to the basal keratinocyte. A recent study involving 50 Rwandan female sex workers found that vaginal dysbiosis correlated with protein markers that were indicative of disruption of the cervicovaginal mucosal barrier, finding evidence of cytoskeletal alterations and increased markers of cell death and proteolysis (Borgdorff et al 2016b). These

changes are a potential route of entry for the HPV virus. The authors also found that increasing VMB diversity was associated with higher levels of pro-inflammatory cytokines, a finding that is supported by the results of other studies (Hedge et al 2006, Jespers et al 2017). Chronic inflammation is a known risk factor for carcinogenesis, providing a possible mechanism by which a dysbiotic VMB could produce conditions that favour the development of cancer. Furthermore, epithelial barrier disruption and inflammation may lead to the activation of cellular repair mechanisms and associated increased cellular division, favouring the replication of the HPV virus. It is possible that the cellular damage associated with a VMB typical of bacterial vaginosis is due to specific bacterial species. The bacterium *G. vaginalis* is often associated with BV and is capable of producing the cellular toxin vaginolysin, which has a cytotoxic effect on human epithelial and cervical cells (Gelber et al 2008), suggesting that it could be directly responsible for the cellular death associated with a high diversity VMB. Vaginolysin is a cholesterol-dependent cytolysin and interestingly the bacterium *L. iners* contains an analoguous protein-encoding gene. Transcription of both vaginolysin and the *L. iners* cytotoxin are upregulated in BV (Macklaim et al 2013).

1.4 The HIV epidemic and HPV in Africa

Infection with human immunodeficiency virus (HIV) and the associated progression to acquired immunodeficiency syndrome (AIDS) remains one of the biggest global health challenges facing humanity today. In 2012, there were an estimated 35.3 million people living with HIV across the globe (UNAIDS 2013). While this represents an increase over previous years, this appears to be a consequence of better treatments and treatment coverage resulting in better survival of HIV positive patients, rather than higher infection rates. Accordingly, the estimated number of new infections was around 2.3 million in 2012, representing a 33% decrease compared with 2001 (UNAIDS 2013). With an estimated 70% of all new infections and an average prevalence of 5.8%, sub-Saharan Africa represents the epicentre of the current HIV crisis. Within that region, the Republic of South Africa has the highest number of people living with HIV, totalling 6.1 million in 2012. In other words, one in six HIV-positive people live in South Africa (UNAIDS 2013).

Sub-Saharan Africa also suffers from high prevalences of other infectious diseases such as malaria, herpes simplex virus type 2 and tuberculosis. Co-infections with HIV and other infectious disease are therefore a common phenomenon, and may be

a factor driving the high HIV transmission rates in this region (Barnabas et al 2011). Additionally, co-infections are responsible for the increased risk of morbidity and mortality in people living with HIV. This includes infection with human papillomaviruses.

Previous studies have shown that women with HIV/AIDS have a higher prevalence of HPV than the general population. In one meta-analysis, the estimated overall prevalence of HPV in women with HIV without cytological anomalies was 36.3%, and within that, the average prevalence was higher in the five studies carried out in sub-Saharan Africa at 57% (Clifford et al 2006). In a recent study conducted in Cape Town, South Africa, the prevalence of HPV infection amongst HIV-positive women was similarly high at 52%, compared with only 21% in the control population (McDonald et al 2014). A further study in Ouagadougou, Burkina Faso found an even higher prevalence of 60% in women with HIV (Djigma et al 2011). This increased prevalence is most likely related to increased persistence of the virus due to HIV-induced immunosuppression (Moscicki et al 2004). Accordingly, in HIV positive women, the prevalence of HPV increases with worsening immune status as measured by a decrease in CD4 count. Interestingly, the association is less strong for HPV-16, perhaps because this type is better able to subvert the immune system, even in immunocompetent individuals (Strickler et al 2003). The increased prevalence of HPV infection in HIV positive women translates into an increased risk of in situ and invasive cervical cancer in patients living with HIV, and this risk increases as HIV progresses (Frisch et al 2000, Goedert et al 1998, Mbulaiteye et al 2003). This is likely to be at least partly due to the fact that one of the main risk factors for the development of serious sequelae following infection with HR-HPV is persistence of the virus (Scheurer et al 2005) and several studies have shown that HPV infection is more likely to persist in women who are co-infected with HIV (Denny et al 2012). As a result, HPV associated cervical cancer is of particular concern in HIV positive women and this group of women are therefore the focus of the present study.

1.5 Cervical Cancer Prevention

1.5.1 Cervical cancer prevention strategies

There are several approaches being employed to prevent morbidity and mortality from cervical cancer worldwide. These approaches can be divided into primary prevention of HPV infection with the use of public education and the administration

of vaccines, and secondary prevention in the form of screening programmes for early identification of women with persistent HR-HPV infection.

Widespread screening for cervical cancer was initiated following the demonstration by Papanicolaou in the 1940s that early HPV-induced alterations could be detected by cytological examination of exfoliated cervical cells allowing the early identification of asymptomatic precancerous changes - a test now widely known as the Pap smear (Vilos 1998). The Pap smear identifies cellular changes that are indicative of precancerous or cancerous histological changes in the cervix. The degree of dyskaryosis (abnormality of the nucleus) observed partly corresponds with the histological grade of cervical neoplasia (Impey and Child 2012). The test has been adapted to try and lower the rate of false negatives with the use of liquid-based cytology media that reduce morphologic artefacts and the use of computer technology to screen slides (Nanda et al 2000). Despite this, it is important to remember that the Pap smear is a screening test that does not have sufficient sensitivity and specificity to be used as a diagnostic test: a negative result does not rule out abnormal cervical changes and a positive result has to be confirmed by histology. The true number of cases that are prevented by screening is difficult to estimate, but studies suggest that the incidence of cervical cancer in countries with effective programmes is reduced by around 50% (Hoppenot et al 2012).

Alternative screening approaches to cervical cytology are available and, although inferior, remain potentially useful in low resource settings as they are more affordable. This includes visual inspection with acetic acid (VIA) and visual inspection with Lugol's iodine (VILI) which involve visual examination of the cervix using a speculum after the application of acetic acid or Lugol's iodine, respectively. In both cases, this application differentially stains abnormal cervical epithelium. The advantages of these tests include the fact that a result is immediately available, removing the need for an effective recall system and existing laboratory infrastructure. However, although both tests have comparable sensitivity to cytology, they suffer from considerably lower specificity and are prone to observer bias, which currently prohibits their widespread use in screening programmes (Sankaranarayanan et al 2012).

Due to the invariable presence of HR-HPV infection at some point in the progression to cervical cancer, adjunctive testing with HR-HPV testing has been

used, either to allow extension of recall intervals or to rule out progressive lesions following an abnormal Pap test result (Cuzick et al 2006, Impey and Child 2012). In some settings, HPV testing is more sensitive than cytology (Cuzick et al 2006) and has been considered as a sole screening test (Wright 2007). However due to the ubiquitous nature of the virus, the specificity of such testing is relatively low, greatly increasing the cost of screening and has therefore not been widely adopted.

Since 2007, many developed countries have additionally begun to offer vaccination against HR-HPV infection. These vaccinations are effective against the two most common and pathogenic HPV types HPV-16 and HPV-18 and some also provide protection against the low risk HPV types 6 and 8 that cause genital warts. A systematic review and meta-analysis of the success of these vaccinations in high-income countries showed that the overall prevalence of HPV-16 and HPV-18 in girls aged between 13 and 19 years dropped by 64%. Furthermore, a protective herd effect was observed in countries where coverage amongst girls was at least 50% (Drolet et al 2015). Recently, the pharmaceutical company Merck has licenced a vaccine that extends coverage to the HR-HPV types 31, 33, 45, 52, and 58 (Kirby 2015). However, there is epidemiological evidence that the quadrivalent vaccine already provides some cross-protection against at least some of these types (Drolet et al 2015) and improved efficacy of this new vaccine has not yet been conclusively shown. Despite the success of HPV vaccination programmes, they do not offer complete protection (and are currently only offered to young girls) and most high income countries continue to employ screening programmes for cervical cancer prevention.

There is currently no epidemiological information on vaccine efficacy available from developing countries, where high vaccine cost has prevented widespread uptake. However, beginning in 2011 some lower-income African countries have initiated HPV vaccination programmes with the aid of supplier-donated vaccines and funding from the Global Alliance for Vaccines and Immunisation (Bustreo et al 2015). Although differences in sexual behaviour, HPV epidemiology and high co-infection rates with diseases such as HIV may alter the efficacy of vaccines, there is currently no evidence to suggest that efficacy would be reduced in low- or middle-income countries (Drolet et al 2015). However, there is some concern that differences in the prevalence of HR-HPV types in women with HIV might result in lower vaccine efficacy in this group (Denny et al 2012). The fact that HPV vaccination has recently

been shown to be safe and immunogenic in HIV-positive women is encouraging (Denny et al 2013), but effectiveness in preventing high grade intraepithelial lesions and cervical cancer in HIV-infected women remains to be demonstrated (Toft et al 2014).

Nationwide screening and vaccination can be very effective at reducing the incidence of cervical cancer, but success is greatly reduced by cost and logistic barriers in low resource settings. Alternatives to currently available strategies are therefore needed in countries such as South Africa, where current control programmes are suboptimal.

1.5.2 Cervical cancer prevention in South Africa

South Africa runs a national programme for cervical cancer prevention which offers Pap smears for women. However, the current programme offers only three smears over a woman's life time from the age of 30, with a long screening interval of ten years. As a result, cases could be diagnosed at the point when treatment is no longer effective. This is especially true for women who are HIV positive and in which disease progression is accelerated. Furthermore, success of the programme has been limited by non-implementation in several parts of the country - particularly the most destitute areas - as well as poor uptake, and relatively high loss to follow-up (Botha and Richter 2015, Laubscher et al 2015).

HPV vaccines have been available privately in South Africa since 2008. However, coverage has remained low with cost and insufficient public awareness having been cited as possible causes. In April 2014, the Department of Health initiated a school-based vaccination programme for girls (Botha and Richter 2015) and in due course studies will be able to assess how effective vaccination is in this country where HIV incidence is exceptionally high. Regardless of their efficacy, vaccination programmes are unlikely to be fully implemented in low-resource settings where they are arguably most needed (Brotman et al 2014).

While there are ongoing projects to identify ways to reduce the cost of cervical cancer screening programmes - including the HPV in Africa Research Partnership study, that forms the parent study to this PhD project - it is unlikely that nationwide screening will be implemented in South Africa because of a lack of adequate resources (Laubscher et al 2015). Due to the poor success of cervical cancer

prevention schemes in countries such as South Africa where resources are stretched, there is an urgent need for novel strategies that are both successful and cost-effective.

1.6 Aims of the Study

Owing to the important barrier function of the human microbiota, manipulation of the VMB is a promising candidate as a preventative treatment for HPV and other sexually-transmitted infections. It is particularly important to develop cost-effective strategies for primary HPV disease prevention in HIV-positive women who are at particularly high risk of developing cervical cancer. Exploration of the association between the VMB, HPV infection and cervical cancer in women living with HIV infection is a first step towards improving health outcomes in this group of women and is the main aim of this project.

Additionally, this knowledge is vital in enabling researchers and medical professionals to fully evaluate the outcomes of clinical interventions that have the potential to alter the composition of the VMB. For such clinical trials, an awareness of the significance of an altered VMB state is a prerequisite for assessing whether a potential treatment could pose hidden risks to the patient, for example by increasing their susceptibility to HPV infection.

CHAPTER 2: Laboratory Methods Validation

Part of the text and results in this chapter (sections 2.6 and 2.7) have been published in:

Gill C, Van de Wijgert JH, Blow F, Darby AC (2016). Evaluation of Lysis Methods for the Extraction of Bacterial DNA for Analysis of the Vaginal Microbiota. PloS One 11:e0163148.

I conceived, designed and carried out the experiments and drafted this manuscript. JHvdW and ACD provided input on conception and design. FB assisted with data analysis. All authors contributed to the interpretation of the data and the content of the manuscript and approved the final version.

2.1 Introduction

In order to determine the association between the vaginal microbiome (VMB), HPV infection and cervical cancer it is necessary to accurately determine the types of bacteria that make up the VMB in each sample. Although characterisation of the VMB by sequencing all DNA in the sample (whole genome shotgun sequencing) is considered the gold standard when it comes to determining true sample composition (and also captures other organisms such as fungi and eukaryotes) this relies on a much higher sequencing effort which would be prohibitively expensive for large numbers of samples. A popular approach to characterise the human

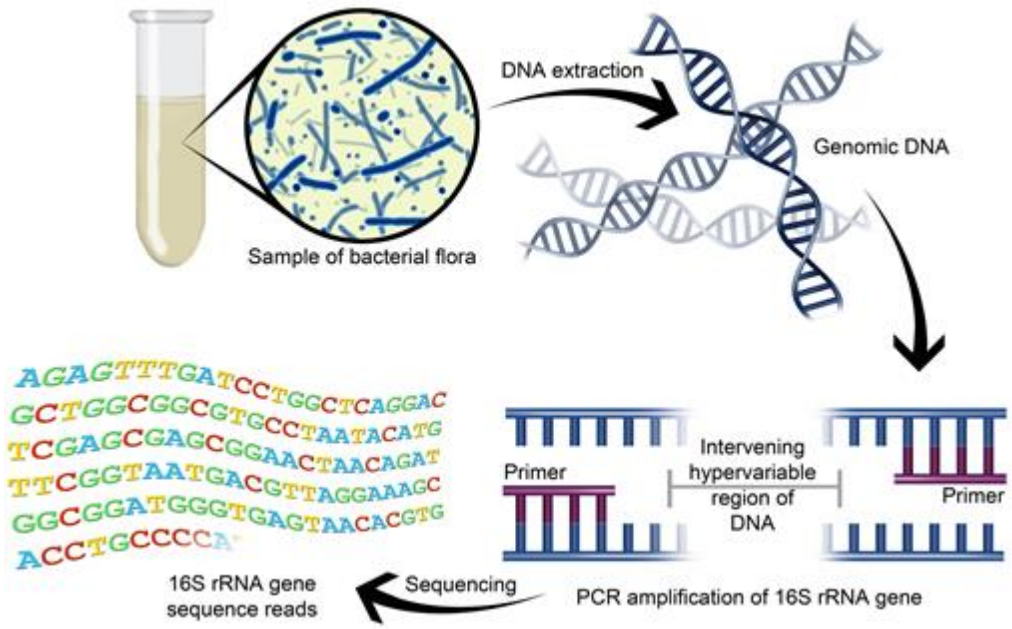


Figure 2.1 Schematic overview of 16S rRNA gene amplicon studies for characterisation of the bacterial microbiota.

microbiota is therefore based on amplifying by PCR and then sequencing a universal bacterial gene, most often a part of the 16S rRNA gene (Figure 2.1). This gene has the advantage that it is present in all bacterial species, in theory resulting in PCR amplicons that are proportionally representative of the bacteria present in the original sample. Additionally, comprehensive 16S rRNA databases are now available, allowing identification of a wide range of bacteria from their 16S rRNA gene sequences (Cole et al 2014, DeSantis et al 2006). For these reasons, 16S rRNA gene sequencing was chosen to characterise the VMB in this study.

Processing of samples for 16S rRNA amplicon studies begins with the collection of suitable samples and their storage under appropriate conditions until they can be processed further. Genomic DNA is then extracted from samples, usually with the use of a commercially available DNA extraction kit. Following this, the 16S rRNA gene is amplified by PCR using primers that are specific to conserved regions of the gene, theoretically creating amplicons from the genomic DNA of all bacteria present in the sample. By design, these amplicons will contain one or more of the 16S rRNA gene's hypervariable regions (which are numbered V1 through to V9) which allows identification of the bacteria whose DNA they were amplified from. Once the amplicons have been created, their DNA can be sequenced using one of several high throughput sequencing platforms. This produces several thousands of DNA sequences for each sample which each represent an amplicon (Figure 2.1). Using computational approaches (known as "bioinformatics"), amplicon sequences can then be assigned to groups based on similarity to other amplicons in the project. The typical cut-off used is 97% similarity, such that sequences within one group share at least 97% of their DNA with the reference or centroid sequence, making them roughly equivalent to bacterial species (see Chapter 3). However, since the definition of bacterial species is complex and based on a variety of factors such as phenotypic characteristics, DNA-DNA hybridisation and/or full length gene sequences (Gevers et al 2005), the resulting groups are not the same as bacterial species and are therefore referred to as "operational taxonomic units" (OTUs) (Nguyen et al 2016). New methods are emerging which do not apply such an arbitrary similarity threshold but instead group reads by other means, increasing the resolution of 16S rRNA studies (see Chapter 3). Ideally, the above steps would result in an accurate profile of the bacterial population in the original sample, giving the proportion of each bacterial OTU present. However, each of these processing steps has the potential to introduce bias, which may result in inaccurate OTU

proportions and even entire groups of bacteria being missed (Brooks et al 2015). Furthermore, PCR amplification has the disadvantage that it may introduce bias by preferential amplification of particular sequences and due to the presence of variable numbers of (and potentially dissimilar) gene copies in different bacterial strains (Nguyen et al 2016). It is vital for the accurate interpretation of results that researchers understand any potential bias and attempt to minimise it. Considerable effort was therefore made to validate DNA storage and extraction for the analysis of vaginal samples, the results of which are presented in this chapter.

2.2 Methods Part I: Sample Storage and DNA Extraction

2.2.1 Sample characteristics

Samples collected for microbiota analysis had been stored in BoonFix® (a patented fixative containing ethanol and polyethylene glycol) at room temperature. This medium has been used for molecular studies of the VMB and was selected for practical reasons prior to the decision to use 16S rRNA amplicon sequencing. Since this medium has not been validated for these types of studies, we set out to determine whether samples stored at room temperature in BoonFix® were suited to microbiome analysis by 16S rRNA gene sequencing. For this purpose, a set of eight paired samples from the endline visit of the HARP study (the parent study of the work presented in the following chapters, see section 4.2) was used. Each sample pair consisted of a vaginal swab sample stored in 2 ml BoonFix® at room temperature and a cervical brush sample taken at the same time and stored in 10 ml PBS-methanol at -80°C. Ethical approval for the determination of the VMB from these samples had been obtained from the local ethics committees at Wits University in Johannesburg, South Africa; the London School of Hygiene and Tropical Medicine, UK; and the University of Liverpool, UK (Physical Interventions Sub-Committee). The purpose of this study was to determine whether differences in microbiome composition could be observed between the different storage and extraction procedures.

2.2.2 DNA extraction

The vaginal swab samples stored in BoonFix® were processed after removal of the swab by thoroughly vortex mixing the remaining liquid and then subjecting 200 µl to 30 min of lysis at 37°C using enzymatic lysis buffer containing lysozyme from chicken egg white (20mg/ml; Sigma-Aldrich, Dorset, UK), followed by DNA extraction using the QIAsymphony DSP Virus/Pathogen Kit (Qiagen, Manchester,

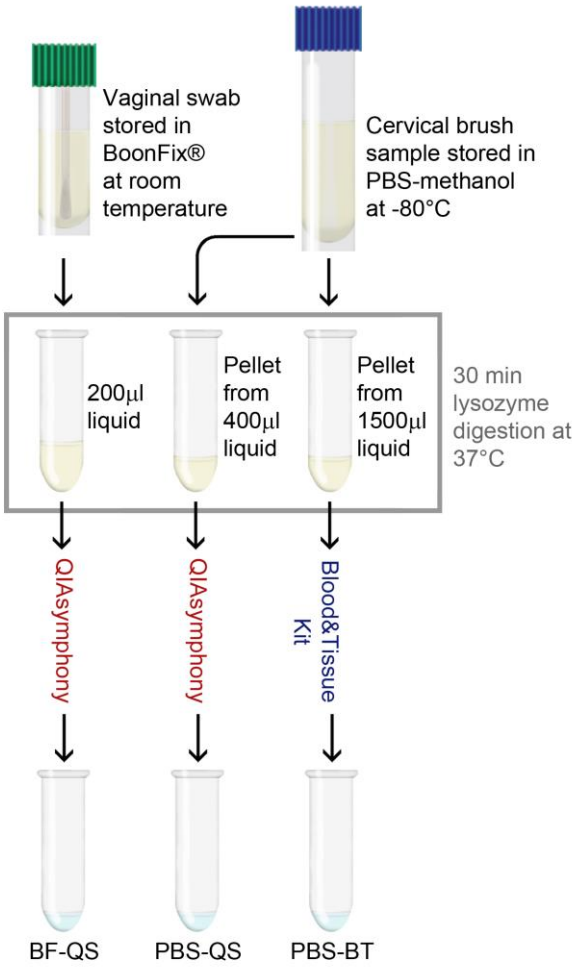


Figure 2.2 Overview of experimental design of sample storage and DNA extraction experiment showing how DNA was extracted from samples.

UK) on the QIAsymphony robot (Complex800_V6_DSP protocol which includes a proteinase K/"buffer AL" digestion step; extract designated "BF-QS", see Figure 2.2). At the same time, two DNA extracts were produced from 400 µl and 1500 µl of the paired PBS-methanol samples as follows: after vortexing, the samples were centrifuged at 5000 x g for 10 minutes and the supernatant was removed. The pellet was then subjected to lysozyme digestion as above, followed by DNA extraction using either the QIAsymphony robot (as above) for the 400 µl sample (extract designated "PBS-QS"), or following the remaining steps in the Qiagen DNeasy Blood and Tissue kit's spin column protocol (equivalent to the recommended pretreatment for Gram-positive bacteria as per the Qiagen DNeasy Blood and Tissue kit Handbook) for the 1500 µl sample (extract designated "PBS-BT"; see Figure 2.2). DNA was eluted in 60 µl of elution buffer for samples run on the QIAsymphony robot and 75 µl of elution buffer for samples extracted using the Qiagen DNeasy Blood and Tissue kit. The genomic DNA concentration of extracts

was determined using the Qubit Fluorometer with the dsDNA HS Assay kit (Invitrogen Life Technologies, Paisley, UK). Negative extraction controls were produced from nuclease free water (Invitrogen, UK) and processed alongside samples using either the QIAsymphony robot (four controls) or the Qiagen DNeasy Blood and Tissue kit (three controls).

2.2.3 Amplicon library preparation

The V3-V4 region of the 16S rRNA gene was amplified in a 50 µl reaction containing no more than 200 ng of genomic DNA (samples over 10 ng/µl were diluted 1:10 to save sample), 25 µl of NEBNext® High-Fidelity 2x PCR Master Mix (New England Biolabs, Hitchin, UK) and 2.5 µl each of a 10 µM concentration of the conserved bacterial 16S rRNA primers 319F 5'-ACTCCTACGGGAGGCAGCAG-3' and 806R 5'-GGACTACHVGGGTWTCTAAT-3' (Fadrosh et al 2014) adapted with linker regions to allow barcoding of sequences using a dual-indexing approach (D'Amore et al 2016). For the first PCR (16S rRNA gene amplification) the samples were initially denatured at 98°C for 30 s, followed by 10 cycles of 98°C for 15 s, 58°C for 15 s and 72°C for 15 s, with a final extension at 72°C for 60 s. The annealing temperature was optimised by gradient PCR on a vaginal sample to provide optimal amplicon yield while allowing universal coverage (higher temperatures produced a narrower band, indicating possible selectivity for some bacterial species over others). The PCR products were then purified using SeraPure magnetic beads (Faircloth and Glenn), before undergoing a second PCR to attach sample-specific barcodes and further amplify the region of interest. The second PCR consisted of a 25 µl reaction containing 10.5 µl of clean PCR product, 12.5 µl of NEBNext® High-Fidelity 2X PCR Master Mix and 1 µl each of a 3 µM concentration of the Illumina specific barcoding primers with the standard Illumina Nextera 8-nt index sequences. Samples were initially denatured at 95°C for 2 min, followed by 15 cycles of 98°C for 20 s, 55°C for 15 s and 72°C for 40 s, with a final extension at 72°C for 60 s. PCR products were purified, eluted in a volume of 15 µl TE buffer (Sigma-Aldrich) and quantified using the Qubit Fluorometer with the dsDNA HS Assay kit to determine amplicon yield. Purified PCR amplicons were run on a 2% agarose gel at 100V to determine purity of the amplicon. Amplicons were then pooled and sequenced on the Illumina MiSeq platform (2x250bp; Illumina, San Diego, CA) at the University of Liverpool Centre for Genomics Research.

2.2.4 Bioinformatics

Sequencing reads were demultiplexed and trimmed for the presence of Illumina adapter sequences and low quality bases (quality threshold Q = 20) using Cutadapt v. 1.2.1 (Martin 2011) and Sickle v. 1.200 (github.com/najoshi/sickle), respectively. The resulting reads were error corrected using SPAdes v 3.1.0 (Bankevich et al 2012) and paired-end alignment was performed using PANDAseq v. 2.4 (Masella et al 2012). The obtained sequences were then binned into OTUs based on 97% sequence similarity using USEARCH v. 5.2.236 (Edgar 2010) through Quantitative Insights Into Microbial Ecology (QIIME v. 1.7.0)(Caporaso et al 2010). Low abundance OTUs were removed (OTUs containing less than 4 reads in total). Taxonomic assignment of representative sequences (most abundant) was carried out for each OTU by RDP classifier against the Greengenes 13_8 database in QIIME and assignments were corrected manually by NCBI BLAST search (Zhang et al 2000) for all the most abundant OTUs (≥1% in at least one extract).

2.2.5 Data analysis

Calculation of alpha and beta diversity measures and statistical analyses were performed in R version 3.2.2 (R Core Team 2015) and using the vegan package version 2.3-1 (Oksanen et al 2015). Observed OTUs and the Simpson Index (1-D) were calculated to assess differences in alpha diversity. Hypothesis testing relating to DNA yield and alpha diversity was performed using the Wilcoxon signed-rank test for paired samples, thereby correcting for differences due to variation between study participants. Bray-Curtis dissimilarity and its complement, Bray-Curtis similarity, were used to report and assess differences in beta diversity. Permutational multivariate ANOVA (PERMANOVA) (Anderson 2001) was used to assess differences in beta diversity between different extracts. The OTU heatmap and the principal coordinate plots were generated in R using the phyloseq package version 1.14.0 (McMurdie and Holmes 2013).

2.3 Results Part I: Sample Storage and DNA Extraction

2.3.1 DNA extraction yield

When comparing the extracts produced on the QIAsymphony robot from the BoonFix® sample (BF-QS) with those from the PBS-methanol sample (PBS-QS), the BF-QS samples had a lower DNA yield (median = 12.7 ng, range 4.4-45.4 ng) when compared to PBS-QS (median = 39.9 ng, range 2.2-104 ng). When comparing the DNA yield from the extraction of the PBS-methanol samples that

were produced using the QIAsymphony robot (PBS-QS) with that produced using the Qiagen DNeasy Blood and Tissue kit (PBS-BT), the PBS-BT samples had a higher yield even after correction for higher sample input volume (median yield from 400 µl of sample = 157 ng, range 20-982 ng). Neither of these differences were statistically significant (Wilcoxon signed rank test, P = 0.11 in both cases). By comparison, the total DNA yield from the negative extraction controls was <1.5 ng (below the measurable range) with the Qiagen DNeasy Blood and Tissue kit and between 13 and 33 ng with the QIAsymphony robot.

2.3.2 Amplicon PCR product

The amplicon yield ranged from <0.1 (recorded as 0) to 16 ng/µl. The median amplicon concentration for each method mirrored the DNA extract concentration, being the lowest for BF-QS (median = 0 ng/µl), higher for PBS-QS (median = 0.51 ng/µl) and highest for PBS-BT (median = 4.98 ng/µl). For those PCR products where amplicon yield was sufficient (DNA concentrations of at least 0.382 ng/µl), gel electrophoresis showed a single band at approximately 600bp, with the exception of a single PBS-BT extract which produced a double-band (see Appendix A). This sample had a relatively high DNA input and produced a correspondingly high amplicon yield (11.9 ng/ µl). Sequencing showed that this sample contained a diverse array of bacteria and the double band could be due to differences in amplicon length of different bacterial species, although non-specific amplification cannot be ruled out. Unsurprisingly, there was a positive correlation between DNA input into the PCR reaction and amplicon yield (Spearman's rank correlation coefficient = 0.71, P <0.0001).

2.3.3 Sequencing results

A total of 4,288,509 paired 16S rRNA sequence reads were generated from the cervical brush and vaginal swab samples (8 sample pairs x 3 extractions). After error correction, paired-end alignment and assignment to OTUs, a total of 2,235,017 reads were retained, with a median read count of 90,506 reads ranging from 1,222 to 169,158 reads. A total of 573 OTUs were identified, of which 51 were present at 1% or more in at least one sample extract. Positive and negative controls were included on the sequencing run. The main contaminant present in the profiles of all the negative DNA extraction controls was a *Rhodanobacter* sp. (66.9-86.6%). This OTU was virtually absent from the negative PCR control (2 reads mapping to this OTU are probably the result of incorrect assignment to this sample) and has

therefore most likely originated from the DNA extraction kits. Although two different extraction kits have been used, they are both made by the same manufacturer and were ordered around the same time. The *Rhodanobacter* OTU was also present in samples at up to 25.8% (present at >1% in 10 extracts; median 0.5%), demonstrating that there was significant reagent contamination in some samples. The proportion of this OTU was found to be inversely proportional to PCR product concentration (see Appendix B), indicating that contamination is more important for low biomass samples, which is consistent with the findings of other studies (Salter et al 2014). OTUs that were identified as reagent contaminants (i.e. those that were proportionally more prevalent in negative extraction controls and where the difference was significant by t test) were removed prior to data analysis. With the exception of *Rhodanobacter* sp. and *Pseudoalteromonas* sp. (the latter being present at 1.6% in one extract which also had a high proportion of *Rhodanobacter* at 20.2%), these were not present in any sample extract at more than 0.2%. Following removal of these contaminants (from here on described as "reagent contaminants" to distinguish them from sample contaminants described later), 379 OTUs remained with a read count ranging from 944 to 169,142 per sample. The positive control sample contained DNA from *Lactobacillus amylovorus* only and this sample was dominated by three OTUs (together making up 96.9% of the sample), each assigned to *Lactobacillus* sp. by RDP classifier and identified further as *Lactobacillus crispatus* group (includes *L. acidophilus*, *L. helveticus*, *L. gallinarum*, *L. crispatus*, *L. jensenii* and *L. delbruekii*) by BLAST search. Although this 'OTU-splitting' could be due to actual biological differences between reads or read error, re-analysis of the positive control at a later date suggested that this was most likely due to an error correction step integrated in the OTU picking step of QIIME (see Appendix C). The remaining OTUs in the positive control were for the most part also identified as *L. crispatus* group, with the two largest OTUs that were not identified as such being *Lactobacillus iners* (0.02%) and *Gardnerella* (0.002%), the most abundant OTUs across the run as a whole and therefore most likely the result of incorrect assignment to this sample (also known as barcode switching), although cross-contamination cannot be ruled out.

2.3.4 Vaginal bacterial community composition

As is typical of the vaginal niche, the bacterial community profiles obtained for each study participant were either low diversity (dominated by either *L. crispatus* group and/or *L. iners*), or high diversity containing a mixture of strict and facultative

anaerobes including *Gardnerella vaginalis*, *Atopobium vaginae*, *Prevotella*, *Sneathia/Leptotrichia*, *Megasphaera* and the bacterial vaginosis associated bacteria BVAB1 and BVAB2 (Figure 2.3). At a glance the PBS-methanol samples extracted on the QIAsymphony look similar to their counterparts extracted with the DNeasy Blood and Tissue kit. Broadly, this also holds true when comparing the BoonFix®

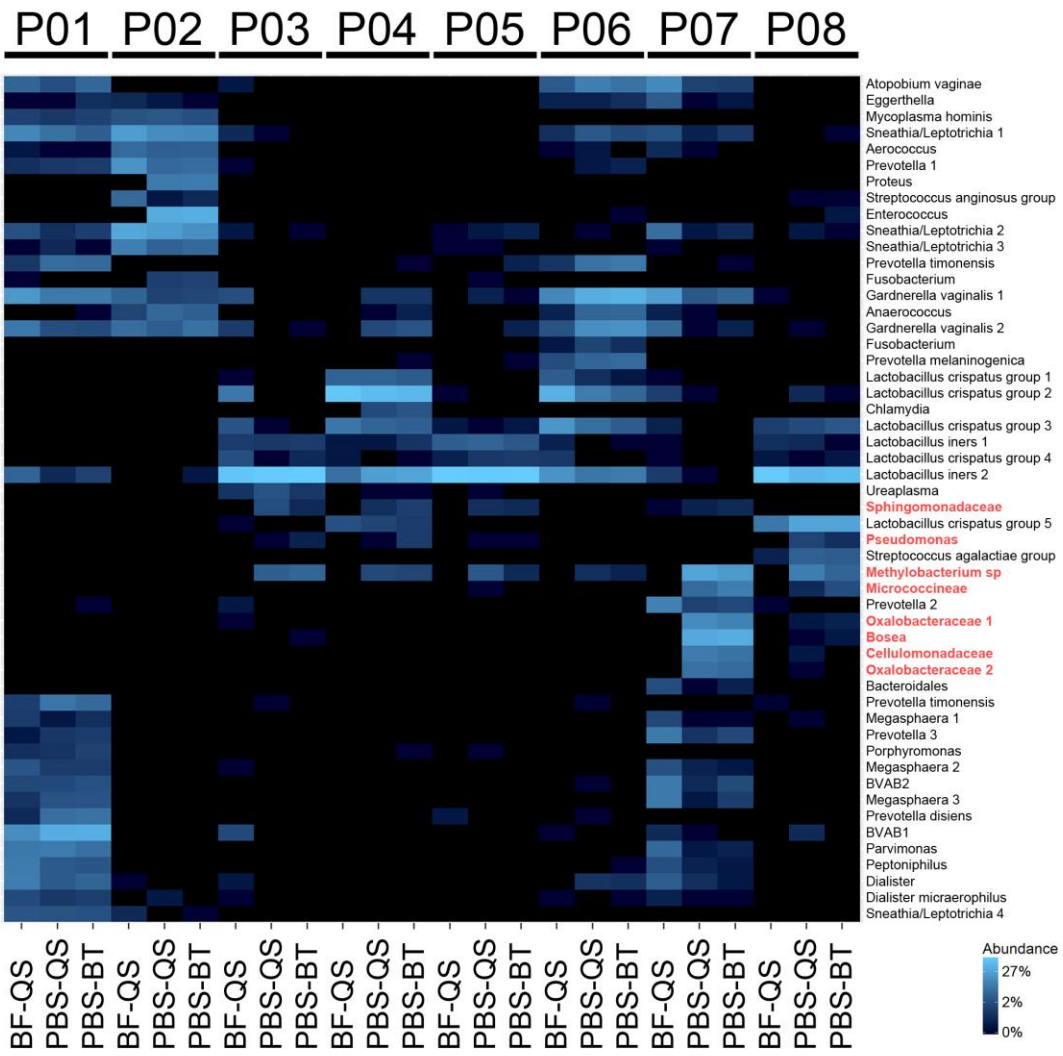


Figure 2.3 Heat map showing most abundant operational taxonomic units in sample storage and DNA extraction experiment. Sample extracts are arranged by hierarchical clustering, after removal of reagent contaminants. All OTUs that were present at 1% or higher in at least one sample extract are shown. Extracts are arranged according to sample, with the participant identifier given at the top and extraction method indicated at the bottom. OTU taxonomy is given on the right, with putative sample contaminants indicated in red. Reads have been assigned to OTUs based on 97% sequence similarity of the V3-V4 region. Note that in some cases this has resulted in multiple OTUs with the same taxonomic species identifier, which may be due to error correction in the QIIME OTU picking step (see Appendix C), intraspecies variability in this region of the gene, or incorrect base calling. *Lactobacillus* species that could not be identified to species level at the 97% cut-off have been assigned to *L. crispatus* group (which includes *L. acidophilus*, *L. helveticus*, *L. gallinarum*, *L. crispatus*, *L. jensenii* and *L. delbruekii*).

with the PBS-methanol samples but there are some striking differences among some sample pairs. This is particularly true in the case of subject P07 (Figure 2.3), in which several OTUs are dominant in both PBS-methanol extracts but completely absent in the BoonFix® sample. These OTUs have been identified as *Methylobacterium* sp., Micrococcineae, *Oxalobacteraceae*, *Bosea* sp, and *Cellulomonadaceae*. *Methylobacterium* spp. are normally associated with soil and water (Dourado et al 2015) and have been identified as contaminants in sequencing experiments previously (Salter et al 2014). This OTU is present in six of the PBS-methanol samples (P03-P08). These bacteria are strict aerobes and are able to grow using methanol (Dourado et al 2015), making it likely that they contaminated the PBS-methanol medium prior to sample collection. In line with that theory, PBS-methanol samples from P01 and P02, which did not have any significant *Methylobacterium* contamination, were collected earlier than the other samples. Three of the other contaminants in P07, namely bacteria in the suborder Micrococcineae, bacteria in the family *Oxalobacteraceae* and *Bosea* sp, have also been reported as contaminants (Salter et al 2014). A further difference is seen between PBS-methanol and BoonFix® extracts of participant P02. In this case, the PBS-methanol samples contain *Proteus* and *Enterococcus*, whereas the BoonFix® extract does not. Both genera are found as commensals in the intestine and sometimes cause opportunistic infections (Chow et al 2011). This difference is interesting as the PBS-methanol sample was taken from the cervix, whereas the BoonFix® sample was obtained from the lateral vaginal wall. Although previous studies have shown that the microbiota profiles from these locations are very similar (Anahtar et al 2015), it could be that these two OTUs represent true colonisation of the cervix, and concurrent absence from the midvagina. It has also been shown that, while generally highly similar, there may be differences in samples obtained using swabs such as were used to collect the BoonFix® sample when compared to cytobrushes such as were used for the collection of the PBS-methanol sample (Mitra et al 2017). Alternatively, contamination of this sample with these bacteria may have occurred at the time of collection. Further analyses were carried out after these contaminants (hereafter referred to as "sample contaminants" to distinguish them from reagent contaminants) were removed (including *Proteus* and *Enterococcus* which may or may not be true contaminants). This resulted in 366 OTUs remaining with a read count ranging from 943 to 169,110 per sample. Samples were rarefied to 943 reads for all further analyses. However, based on rarefaction curves of Faith's phylogenetic diversity index, this rarefaction depth was borderline in terms of being able to assess a sample's full diversity and all analyses

were repeated using a depth of 4901 reads (the next smallest read count), but this did not alter the statistical significance of any of the comparisons described below.

2.3.5 Effect on observed alpha diversity

The presence or absence of OTUs was consistent between the two different extraction methods (i.e. between extract PBS-QS made using the QIAsymphony robot and extract PBS-BT using the Qiagen DNeasy Blood and Tissue kit) and any differences were due to low abundance OTUs that made up no more than 0.4% of any extract. The picture was more complicated when comparing the PBS-methanol extract (PBS-QS) with the BoonFix® extract (BF-QS) produced using the QIAsymphony robot, due to the presence of the sample contaminants listed above (see Figure 2.3). However, after these were removed there was overall agreement between sample pairs with the largest relative abundance of any discordant OTU being 1.8%. There was no statistically significant difference in the number of observed OTUs between the PBS-QS and PBS-BT extracts (Wilcoxon signed rank test; P = 0.29) or between the PBS-QS and BF-QS extracts (P = 0.20). Calculation of the Simpson Index (1-D) confirmed a wide range of diversities (range = 0.11-0.92). The degree of variation between the PBS-QS and PBS-BT extracts was small

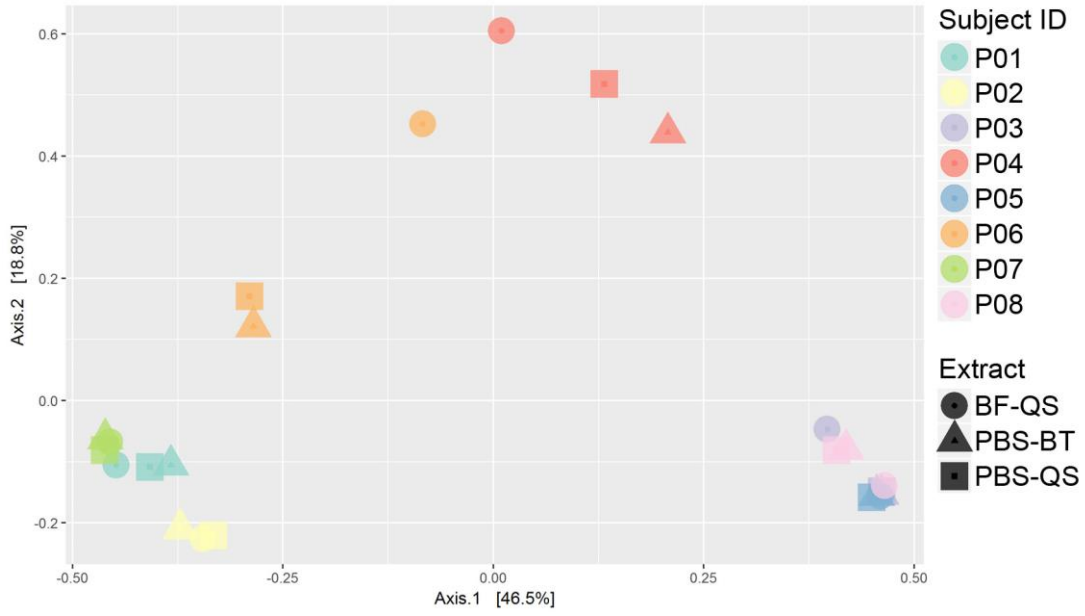


Figure 2.4 Principal coordinate analysis ordination of a Bray-Curtis dissimilarity matrix for the sample storage and DNA extraction experiment. Bray-Curtis dissimilarity was performed on read counts after removal of reagent and sample contaminants and rarefaction to 943 reads. Extracts are coloured by subject of origin and the sample type and extraction method are indicated by different shapes (see key). Extracts cluster closely with other extracts originating from the same subject, with the BF-QS extracts being slightly further removed from the other two extracts.

(maximum difference between extracts = 0.06), but was considerably higher between the BF-QS and PBS-QS extracts (maximum difference between extracts was seen for sample P08 = 0.39). However, these differences were not statistically significant (Wilcoxon signed rank test; P = 0.38, P = 0.48, respectively). Due to the OTU splitting described above, the monoculture positive control had a higher Simpson index than expected (0.56).

2.3.6 Effect on observed beta diversity

Between-extract diversity was calculated using Bray-Curtis similarity. Within-subject similarity ranged from 81.5-98.1% between the two PBS-methanol extracts and from 47.1-95.3% between the PBS-methanol and BoonFix® extracts. In both cases, any differences appeared to be mainly due to differences in OTU relative abundance, rather than differences in presence or absence of OTUs.

PERMANOVA analysis of Bray-Curtis dissimilarity showed that the differences between extracts originating from different women (R2 = 0.92, P = 0.001) were far greater than differences between different extraction methods (R2 = 0.005, P = 0.40) or between sample types (R2 = 0.006, P = 0.28). Therefore, although this study may have lacked power to find a difference between the PBS-methanol and BoonFix® extracts or between the two extraction methods, the effect of either is likely to be much smaller than the differences caused by inter-subject variation. This is reflected in the clustering of extracts by principal coordinate analysis ordination of the Bray-Curtis dissimilarity matrix (Figure 2.4), which resulted in clustering of the extracts by sample rather than lysis method or sample type.

2.4 Methods Part II: Yield Optimisation

2.4.1 Sample characteristics

As the results above showed that the DNA yield from some study samples was low and resulted in proportionally higher levels of reagent contamination an attempt was made to optimise the DNA yield from vaginal swab samples stored at room temperature in BoonFix®. To do so, a set of ten samples from the HARP study (see section 4.2) was randomly selected among those that did not meet the eligibility criteria for the main study described later (see Chapter 4) and for which sufficient material was available. Ethical approval for determination of the VMB from these samples was obtained as above (see section 2.2.1).

2.4.2 DNA extraction

Samples were thoroughly mixed by vortexing and divided into aliquots for DNA extraction using six different methods as follows (Figure 2.5):

Three extracts were produced by subjecting the swab head and 166 µl of liquid each to 30 min of lysis at 37°C using enzymatic lysis buffer containing lysozyme from chicken egg white (20mg/ml; Sigma-Aldrich, Dorset, UK). Following this, one extract (designated "S-QS") was produced by extracting the resulting liquid using the QIAsymphony DSP Virus/Pathogen Kit (Qiagen, Manchester, UK) on the QIAsymphony robot (Complex800_V6_DSP protocol which includes a proteinase K/"buffer AL" digestion step). Proteinase K and Buffer AL (Qiagen) were added to

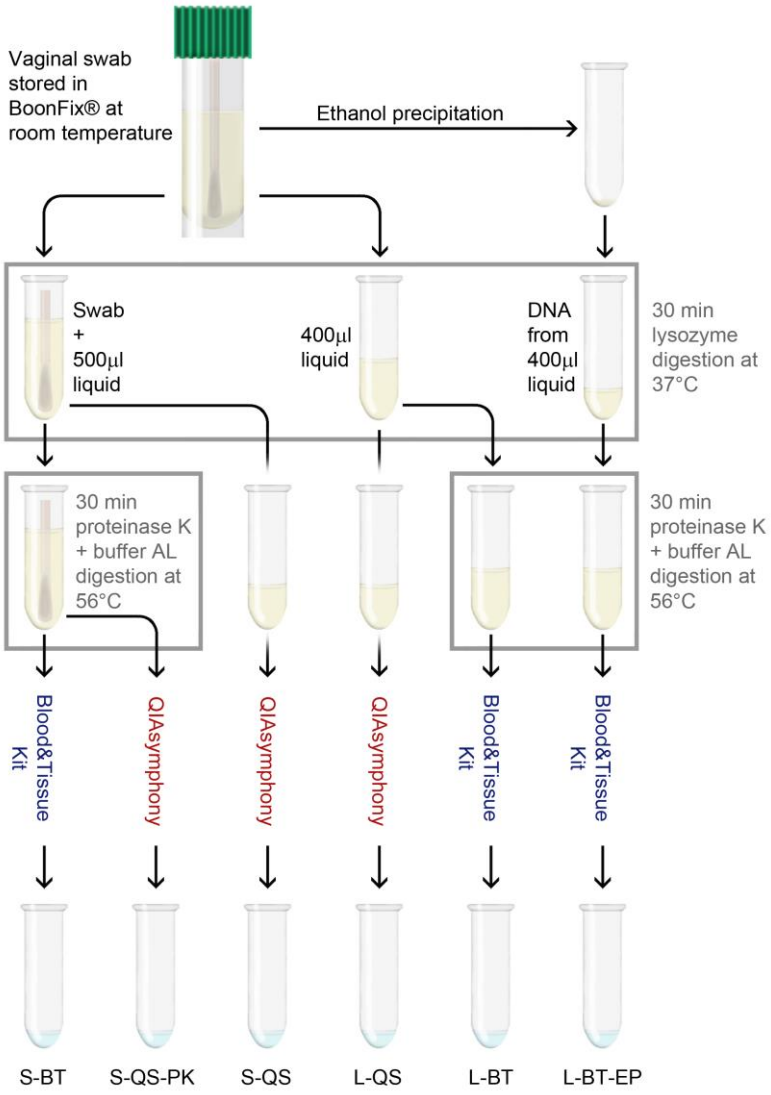


Figure 2.5 Overview of experimental design for yield optimisation from BoonFix®-stored samples. Schematic showing how DNA was extracted from samples.

the remaining liquid and swab head, and incubated at 56°C for 30 min. After discarding the swab, half of the sample was extracted using the QIAsymphony DSP Virus/Pathogen Kit on the QIAsymphony robot (thereby undergoing a second proteinase K/"buffer AL" digestion without the swab head; extract designated "S-QS-PK") and the other half was extracted following the remaining steps in the Qiagen DNeasy Blood and Tissue kit's spin column protocol (equivalent to the recommended pretreatment for Gram-positive bacteria as per the Qiagen DNeasy Blood and Tissue kit Handbook; extract designated "S-BT").Additionally, three extracts were produced from the liquid portion of the BoonFix® sample without including the swab head in the extraction. Two of these were produced from 200 µl liquid sample each, subjected to 30 min of lysis at 37°C using enzymatic lysis buffer containing lysozyme and then extracted using either the QIAsymphony DSP Virus/Pathogen Kit on the QIAsymphony robot (extract designated "L-QS") or by incubation with Proteinase K and Buffer AL at 56°C for 30 min, followed by the remaining steps in the Qiagen DNeasy Blood and Tissue kit's spin column protocol (equivalent to the recommended pretreatment for Gram-positive bacteria as per the Qiagen DNeasy Blood and Tissue kit Handbook; extract designated "L-BT"). In order to test whether the presence of ethanol in the sample affected the extraction, a third extract was produced from 400 µl of liquid BoonFix® sample by first removing ethanol as follows: pure ethanol (Sigma-Aldrich) and sodium acetate (3M, pH 5.2) were added to the sample to produce a solution containing 70% v/v ethanol (taking account of the ethanol already in the sample) and 3% v/v sodium acetate. This was incubated on ice for 30 minutes and then centrifuged at 14,000 x g for 30 minutes at 4°C. The supernatant was discarded, 1 ml of 70% ethanol added and centrifuged at 16,100 x g for 2 min. The resulting pellet was then air dried before lysis at 37°C for 30 min in enzymatic lysis buffer containing lysozyme, followed by incubation with Proteinase K and Buffer AL at 56°C for 30 min and extraction according to the remaining steps in the Qiagen DNeasy Blood and Tissue kit's spin column protocol (extract designated "L-BT-EP"). DNA was eluted in 60 µl of elution buffer for samples run on the QIAsymphony robot and 75 µl of elution buffer for samples extracted using the Qiagen DNeasy Blood and Tissue kit. The genomic DNA concentration of extracts was determined using the Qubit Fluorometer with the dsDNA HS Assay kit (Invitrogen Life Technologies, Paisley, UK). DNA quality was assessed using the 260/280 optical density ratio measured on the NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, USA), with a measurement of 1.8 and above considered pure. Negative extraction controls were produced from nuclease free water (Invitrogen, UK) and processed alongside samples using either

the QIAsymphony robot (four controls) or the Qiagen DNeasy Blood and Tissue kit (four controls).

2.4.3 Amplicon library preparation

The V3-V4 region of the 16S rRNA gene was amplified as described in section 2.2.3. Purified PCR amplicons were run on a 2% agarose gel at 100V to determine purity of the amplicon.

2.4.4 Data analysis

All statistical testing and graphing of data was performed in R version 3.3.2 (R Core Team 2015). Hypothesis testing relating to DNA yield was performed using the Skillings-Mack Test (Skillings and Mack 1981) in the Skillings.Mack package v1.10, a non-parametric equivalent of the repeated measures ANOVA (which can be performed on experiments with block designs that have missing values), hence correcting for differences in yield occurring due to the sample being extracted. Significant results were followed by pairwise comparisons using the Wilcoxon signed-rank test (with p-values adjusted using the Holm–Bonferroni method). Spearman's rank correlation coefficient was calculated to determine whether there was an association between extract DNA concentration and amplicon yield. Linear regression was used to determine whether this relationship was affected by DNA extract dilution (performed on high yield samples to avoid non-specific amplification) or extraction method.

2.5 Results Part II: Yield Optimisation

2.5.1 DNA extraction yield and purity

The DNA quality and yield using six different methods for extraction of bacterial DNA from vaginal swab samples stored in Boonfix® was compared in this study (Figure 2.6). Methods L-QS and S-QS-PK did not produce any DNA extract with samples F and G, respectively. The reason for this is unknown as no error was reported by the QIAsymphony robot.

Among the different methods tested, the mean total DNA yield was highest for samples extracted using method S-BT (inclusion of the swab head and extraction using the Qiagen DNeasy Blood and Tissue kit) at 1243 ng, followed by method S-QS (extraction using the QIAsymphony robot with inclusion of the swab head in both

prelysis with lysozyme *and* proteinase K/"buffer AL") at 250 ng. All other methods had a much lower mean total yield ranging from 15 to 26 ng. By comparison, total DNA yield from negative extraction controls was <1.5 ng (below the measurable range) with the Qiagen DNeasy Blood and Tissue kit and between 10 and 13 ng with the QIAsymphony robot. It should be noted that the sample input volume was unequal between methods, as this was 400 µl for method L-BT-EP, 200 µl for other methods using only sample liquid and 166 µl for methods including the swab head (plus any swab-associated liquid). This was not adjusted for in the statistical analysis because the volume of liquid and associated biomass retained in the swab could not be accurately estimated, therefore the results reflect the methodology as well as the associated sample volume used.

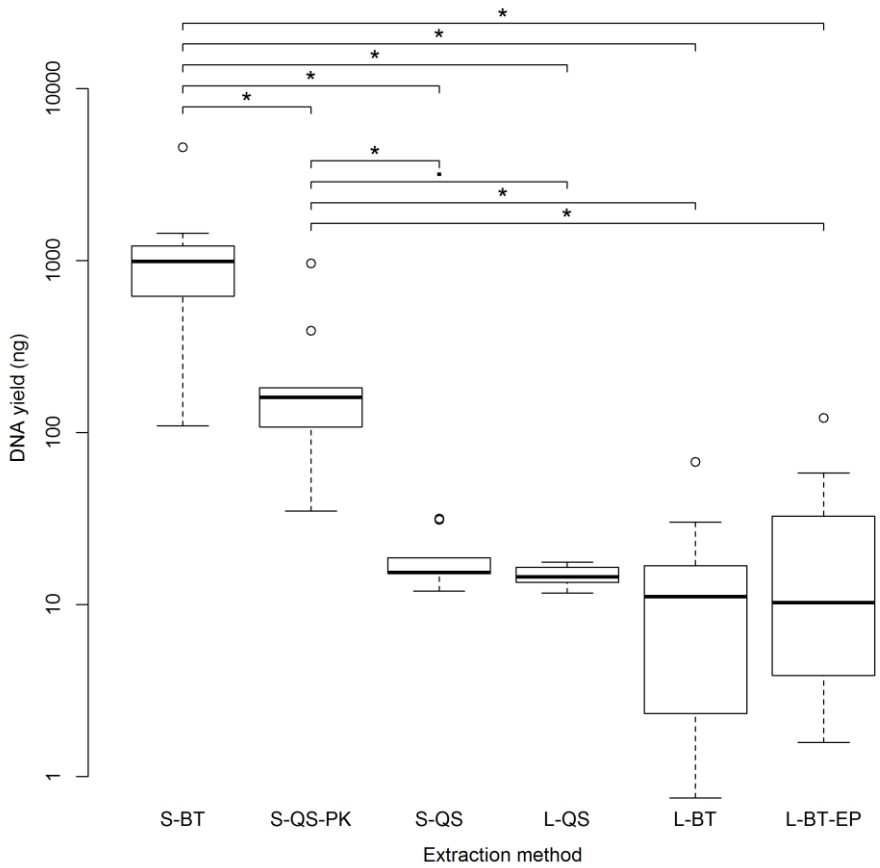


Figure 2.6 Box and whisker plot of DNA yield obtained for samples stored in BoonFix® with each extraction method. Note that the y axis is logarithmic. Boxes extend from the lower quartiles to the upper quartiles with median values indicated by the line within each box. Whiskers represent maximum and minimum values, excluding any outliers (values indicated by circles which lie outside 1.5 times the interquartile range). Significant and near significant differences (paired Wilcoxon signed rank test with Holm–Bonferroni correction) between methods are indicated (•P <0.1; * P <0.05; ** P ≤0.01; *** P ≤0.001).

The Skillings-Mack Test showed that differences between methods were statistically significant (P < 0.0001). Pairwise comparisons showed that method S-BT produced significantly higher DNA yield than all other methods (P = 0.047 when compared with L-QS and S-QS-PK and P = 0.029 compared with all other methods). Note that the difference in P value is due to the failure of an extraction each with methods L-QS and S-QS-P, resulting in reduced statistical power. Excluding method S-BT, method S-QS-PK produced significantly higher DNA yields than all other methods (P = 0.047), except when compared to L-QS (P = 0.055). The latter comparison lacked statistical power since only 8 sample pairs could be compared due to one missing sample from each of the two methods. Therefore the difference is not statistically significant (after correction for multiple testing), despite all S-QS-PK extracts producing considerably higher DNA yields compared to method L-QS (see Figure 2.6).

DNA purity as defined by the 260/280 spectrophotometry absorbance ratio was poor for methods L-BT and L-BT-EP (median of 1.65 and 1.70, respectively) and good for method S-BT (median 2.11). The readings for the methods using the QiaSymphony were generally higher (S-QS, L-QS and S-QS-PK having a median ratio of 3.59, 3.60 and 3.09, respectively and this was due to a higher absorption at 260 nm, rather than lower absorption at 280 nm. This is therefore most likely due to the addition of carrier RNA (which, like DNA, has an absorbance maximum at 260 nm) during the QiaSymphony extraction process since these samples do not contain higher amounts of DNA.

2.5.2 Amplicon PCR optimisation

The amplicon yield ranged from <0.1 (recorded as 0) to 2.5 ng/µl with a median of 0.6 ng/µl and a positively skewed distribution (mirroring DNA concentration of the extracts). One sample and one of the extraction methods (L-QS) failed to produce any measurable amplicon in all cases. For those PCR products where amplicon yield was sufficient, gel electrophoresis showed a single band at approximately 600bp, indicating that the majority of DNA in the sample was amplicon. As expected, there was a positive correlation between the amount of input DNA used in the PCR reaction and amplicon yield (Spearman's rank correlation coefficient = 0.71, P <0.0001; see Figure 2.7). Using a linear regression model with DNA input and whether the sample has been diluted prior to PCR or not (samples with concentrations of 10 ng/µl, N = 7; see section 2.3.3) as fixed effects and sample

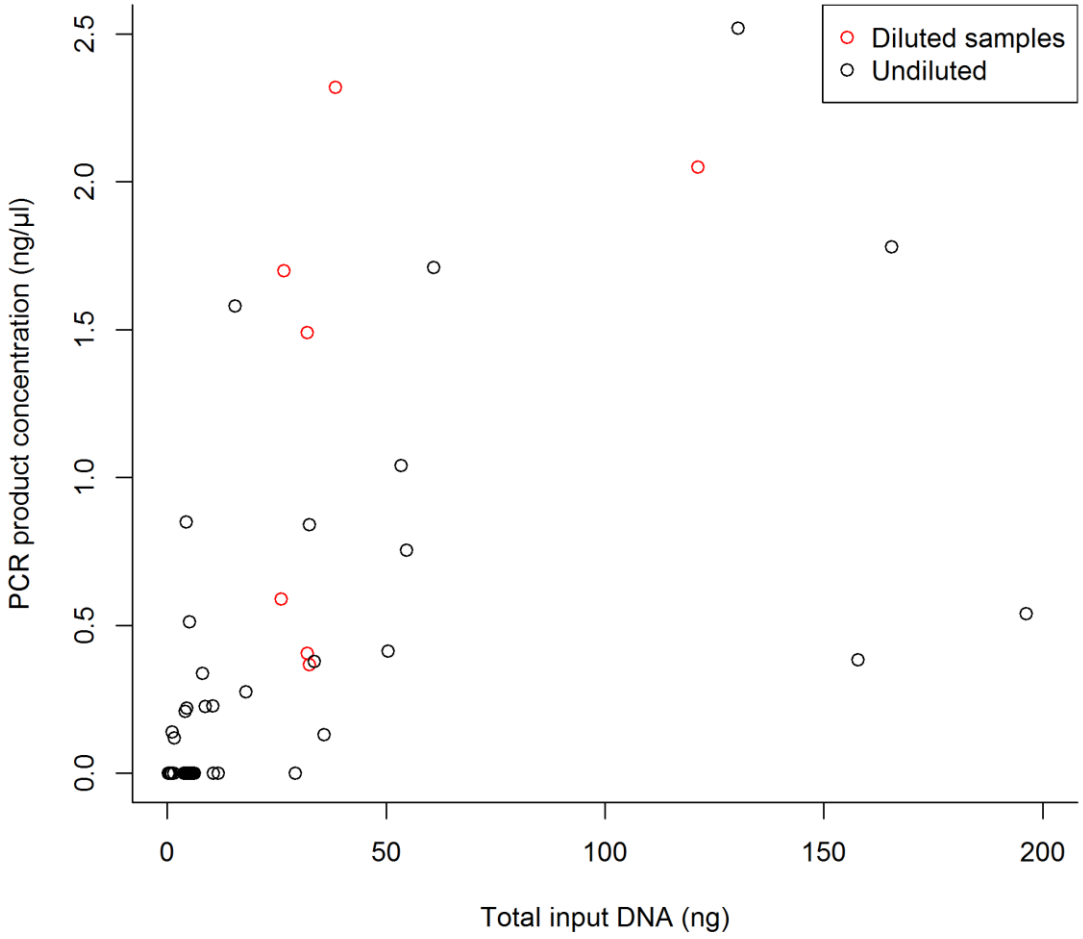


Figure 2.7 Scatter plot of PCR product concentration obtained after two stage PCR against total DNA input for the yield optimisation experiment.

as a random effect (to control for any effect caused by sample and therefore differences in eukaryotic:bacterial DNA), the association between DNA extract concentration and amplicon yield remained highly significant (estimate = 0.0080, P < 0.0001), with dilution also having a positive effect on amplicon yield (estimate = 0.76, P = 0.0001). Addition of extraction method to the model did not improve fit (according to the Akaike information criterion).

2.6 Methods Part III: Bacterial Cell Lysis and Storage in BoonFix®

Studies on the vaginal microbiota most commonly use a commercially available DNA extraction kit (Aagaard et al 2012, Gajer et al 2012, Ravel et al 2011, Shipitsyna et al 2013, Srinivasan et al 2012) but these methods have been poorly validated for studies on the human microbiota, and the choice of kit is often arbitrary. Commercial kits use a combination of different techniques to lyse cells, including mechanical (usually bead beating), chemical and enzymatic lysis and

heating. Methods that include a bead beating step have the advantage that they concurrently homogenise the sample, but this can shear the DNA into short fragments and may increase the risk of contamination during processing (Abusleme et al 2014, Salonen et al 2010). Methods using chemical and enzymatic lysis are less likely to damage DNA, but are thought to increase the potential for extraction bias (Salonen et al 2010). In order to determine whether different pretreatment lysis methods result in significant differences in DNA yield, observed taxa and community structure, we used natural vaginal bacterial communities sampled by cervicovaginal lavage. Additionally, an aliquot from each sample was stored in Boonfix® at room temperature prior to extraction, in order to determine if this provided equivalent results.

2.6.1 Sample characteristics

The 18 cervicovaginal lavage samples used here were a subset of anonymised samples that had been collected in Rwanda as part of a study that aimed to determine whether there was an association between the type of vaginal bacterial community and prevalent infection with sexually transmitted viral diseases (Borgdorff et al 2014). Ethical approval was obtained from the Rwanda National Ethics Committee and the Columbia University Medical Centre Review Board. The purpose of the current study was to evaluate lysis procedures, and samples were chosen solely because the bacterial communities had previously been well-characterised by microarray analysis. We did not have access to personal identifiers and did not use any other data from the study. The 18 samples were chosen to be representative of the community clusters identified previously, including both low diversity communities dominated by either *L. crispatus* or *L. iners* and high diversity communities containing a mixture of strict and facultative anaerobes, representative of the complexity and richness of real vaginal communities. Samples were stored at -80°C until analysis.

2.6.2 Lysis methods

To test for differences in the results of microbiota analyses resulting from different pretreatment lysis strategies, samples were thoroughly mixed by vortexing before dividing into 5 aliquots of 100 µl each and then processed using one of four different lysis protocols (Figure 2.8). Vaginal samples may contain viscous mucoid material and if this was the case, any such material was discarded prior to vortex mixing. Two aliquots (designated "LN1" and "LN2") were subjected to 30 min of lysis at

37°C using enzymatic lysis buffer containing lysozyme from chicken egg white (20mg/ml; Sigma-Aldrich, Dorset, UK). This corresponds to the recommended pretreatment for Gram-positive bacteria as per the Qiagen DNeasy Blood and Tissue kit Handbook (Qiagen, Manchester, UK). One aliquot (designated "LON") was subjected to 16 hours of extended lysis at 37°C using enzymatic lysis buffer containing lysozyme (20 mg/ml). One aliquot (designated "EC") was subjected to 60 min of lysis at 37°C using enzymatic lysis buffer containing lysozyme (20mg/ml), mutanolysin (250U/ml; Sigma-Aldrich) and lysostaphin (22 U/ml; Sigma-Aldrich). The last aliquot (designated "LTL") was subjected to 30 min of lysis at 37°C using enzymatic lysis buffer containing lysozyme (20 mg/ml), followed by 30 s mechanical lysis at 25 Hz using 200 mg of 0.1-mm-diameter zirconia/silica beads in the Tissue Lyser II (Qiagen, Manchester, UK).

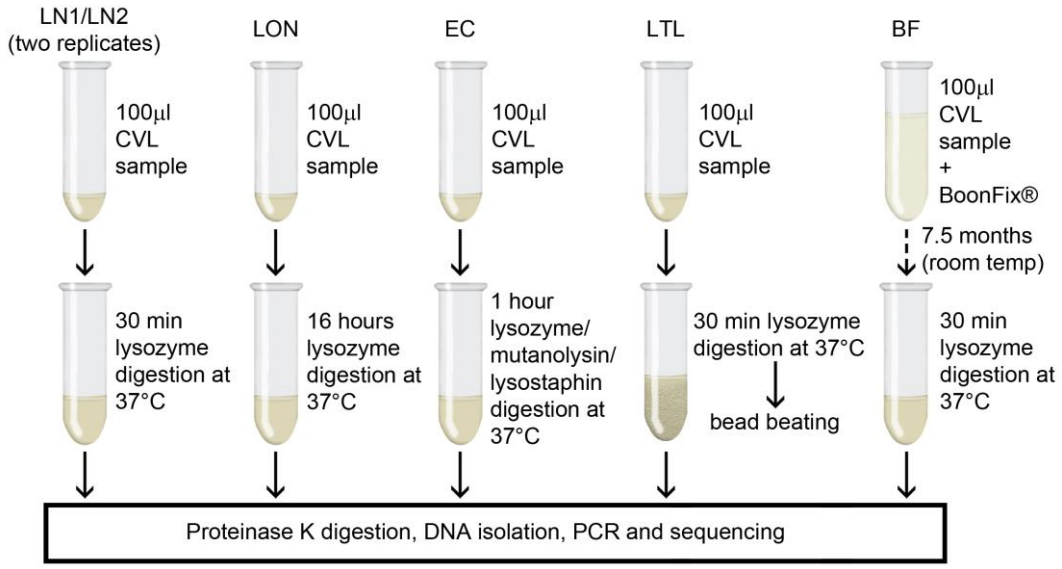


Figure 2.8 Overview of experimental design for cell lysis and BoonFix® storage experiment. Schematic showing how samples were processed for 16S rRNA amplicon sequencing.

2.6.3 DNA extraction

Proteinase K and Buffer AL from the Qiagen DNeasy Blood and Tissue kit (Qiagen) were added to all aliquots before incubation at 56°C for 30 min which was followed by the remaining steps in the kit's spin column protocol, in accordance with the manufacturer's instructions and DNA was eluted in 75 µl of elution buffer. Lysis and DNA extraction was completed for all aliquots within a period of 36 hours using a previously unopened extraction kit and all work was carried out by the same person. The genomic DNA concentration of extracts was determined using the Qubit

Fluorometer with the dsDNA HS Assay kit (Invitrogen Life Technologies, Paisley, UK).

2.6.4 Amplicon library preparation and DNA sequencing

The V3-V4 region of the 16S rRNA gene was amplified in a 25 µl reaction containing 10 ng of genomic DNA, 12.5 µl of NEBNext® High-Fidelity 2x PCR Master Mix and 1.25 µl each of a 10 µM concentration of the conserved bacterial 16S rRNA primers 319F 5'-ACTCCTACGGGAGGCAGCAG-3' and 806R 5'-GGACTACHVGGGTWTCTAAT-3' (Fadrosh et al 2014) adapted with linker regions to allow barcoding of sequences using a dual-indexing approach (D'Amore et al 2016). Thermocycling conditions, barcoding, amplicon purification, DNA quantification and sequencing were performed as described above (see section 2.2.3).

2.6.5 Room temperature storage in BoonFix®

To test the suitability of samples stored at room temperature in BoonFix® fixative for use in microbiome analysis, a further 100 µl aliquot (designated "BF") from each of the 18 samples described in section 2.6.1 above was added to 500 µl of BoonFix® and stored at room temperature for a period of 33 weeks (7.5 months). Samples were then centrifuged at 16,100 x g for 10 min and the supernatant was removed prior to the addition of enzymatic lysis buffer containing lysozyme. Samples were further processed as described for methods "LN1" and "LN2" in section 2.6.2 and DNA was extracted as described in section 2.6.3 (except using a different Qiagen DNeasy Blood and Tissue kit). Amplicon library preparation and sequencing were carried out as described in section 2.6.3 (at a later date and therefore separately to the samples described in 2.6.2 above).

2.6.6 Bioinformatics

Sequencing reads were demultiplexed and primer sequences were trimmed using Cutadapt v. 1.2.1 (Martin 2011). The resulting reads were error corrected using SPAdes v 3.1.0 (Bankevich et al 2012) and paired-end alignment was performed using PEAR v0.9.6 (Zhang et al 2014). The obtained sequences were then binned into *de novo* OTUs based on 97% sequence similarity using USEARCH v. 6.1.544 (Edgar 2010) through Quantitative Insights Into Microbial Ecology (QIIME v. 1.8.0)(Caporaso et al 2010). Taxonomic assignment of representative sequences (most abundant) was carried out for each OTU by RDP classifier against the

Greengenes 13_8 database in QIIME and assignments were checked manually by NCBI BLAST search (Zhang et al 2000) for all the most abundant OTUs (≥1% in at least one extract). OTUs that contained less than 0.005% of total reads were removed as likely sequencing errors (Bokulich et al 2013). Note that this analysis had to be repeated once the sequencing data for the aliquots stored in BoonFix® fixative became available and the results therefore differ slightly from those published in Gill et al (2016), but this has not altered any of the conclusions made. However, the bioinformatics pipeline had to be altered from that originally used, because the second set of sequencing results were of lower quality towards the end of read 2, when compared to the previous results which affected the relative abundance of OTUs (see Appendix D).

2.6.7 Data analysis

Calculation of alpha and beta diversity measures, hierarchical clustering and statistical analyses were performed in R version 3.2.2 (R Core Team 2015) and using the vegan package version 2.3-1 (Oksanen et al 2015). Observed OTUs and the Simpson Index (1-D) were calculated to assess differences in alpha diversity. Hypothesis testing relating to DNA yield and alpha diversity was performed using repeated measures analysis of variance (ANOVA), correcting for differences due to the sample being extracted. Bray-Curtis dissimilarity and its complement, Bray-Curtis similarity, were used to report and assess differences in beta diversity. Permutational multivariate ANOVA (PERMANOVA) (Anderson 2001) was used to assess differences in beta diversity between different lysis methods. Hierarchical clustering was performed using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) on the Bray-Curtis dissimilarity matrix. The OTU heatmap and the principal coordinate plot were generated in R version 3.2.2 using the phyloseq package version 1.14.0 (McMurdie and Holmes 2013).

2.7 Results Part III: Bacterial Cell Lysis and Storage in BoonFix®

A total of 10,374,312 16S rRNA sequence reads were obtained from the 90 cervicovaginal lavage sample extracts (18 samples x 5 extractions) and an additional 2,795,096 reads were obtained from the aliquots stored in BoonFix®. After error correction, paired-end alignment, assignment to OTUs and removal of low abundance OTUs, a total of 9,413,508 reads were retained, with a mean read count of 87,162 reads per sample ranging from 5,307 to 145,306 reads. The lowest read count was found to be sufficient to accurately describe sample diversity based

on rarefaction curves of Faith's phylogenetic diversity index, and unless otherwise stated all further analyses were carried out after samples were rarefied to 5,307 reads. A total of 117 OTUs were identified, of which 41 were present at 1% or more in at least one sample extract. Positive and negative controls were included on both sequencing runs. The main contaminant present in the profiles of all the negative DNA extraction controls was a *Rhodanobacter* sp. (14.9-96.7%). This OTU was absent from the two negative PCR controls and has therefore most likely originated from the DNA extraction kit. Abundance of this OTU was less than or equal to 0.04% of reads of any one sample extract, indicating that contaminants originating from the extraction and amplification process were negligible in this study. The two positive control samples contained DNA from *L. amylovorus* only and were dominated by a single OTU with taxonomical assignment to a *Lactobacillus* sp., making up 99.9% and 100.0% after removal of small OTUs (see section 2.6.6), indicating that the size cut-off used was adequate for removal of erroneous sequences. The next largest OTU in both control samples was also identified as a *Lactobacillus* sp. (0.01% and 0.04%). The representative sequence for this OTU had 97% sequence similarity with the dominant OTU. Interestingly, the presence of these OTUs appeared correlated across the samples, but there were differences in the ratios between the two (which were consistent across different extracts of the same sample). It is possible that this represents biological variation in the sequences of the 16S rRNA gene copies present within different bacterial strains or the presence of different (but closely related) bacterial species. One of the positive controls also contained *Rhodanobacter* (0.01%), and the remaining OTUs present consisted of various vaginal bacteria and the largest of these was *L. iners* (0.02%), the most abundant OTU across all samples. These could represent contamination, but are more likely the result of incorrect sample assignment.

2.7.1 Effect on DNA yield

Four different methods for the pretreatment lysis of bacterial cells in 18 cervicovaginal lavage samples from different women were used in this study. Additionally, long term storage in BoonFix® medium at room temperature was tested (Figure 2.8). Following extraction of DNA using a commercial kit, the total yield of genomic DNA was determined and compared between different lysis methods. The mean DNA yield was highest for aliquots extracted following storage in BoonFix® (method BF, median yield = 64 ng/µl; extracted using pretreatment lysis with lysozyme only) and lowest for samples extracted using enzymatic lysis

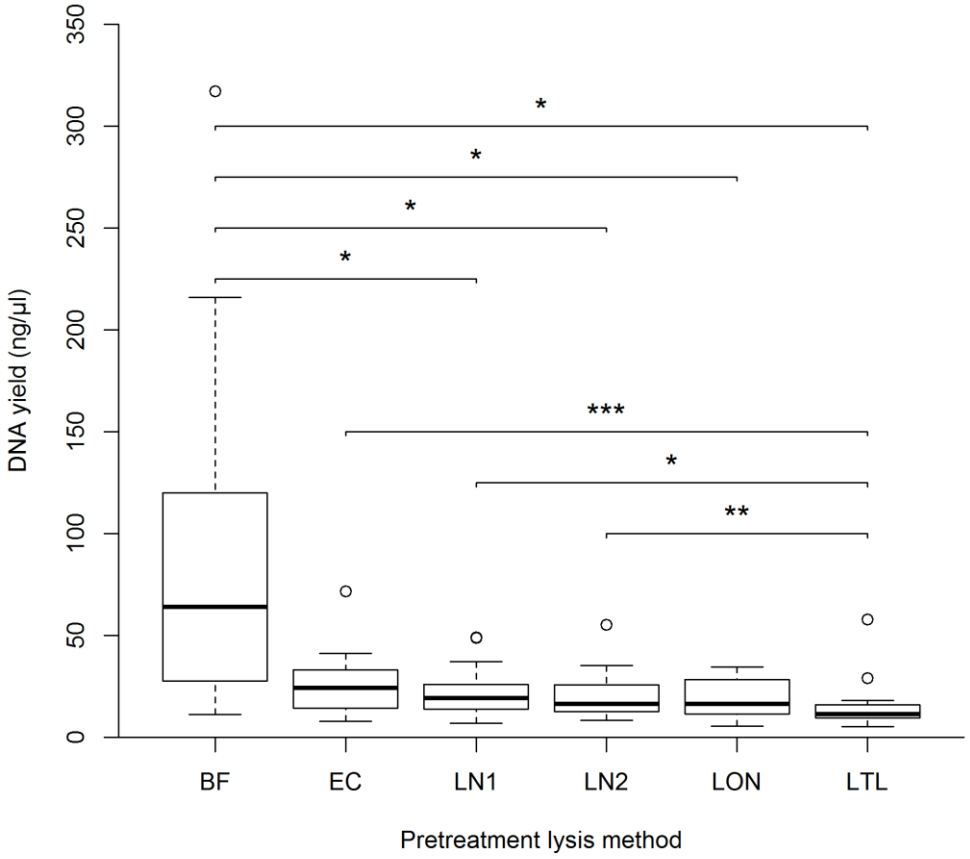


Figure 2.9 Box and whisker plot of DNA yield obtained for samples stored in BoonFix®, and with each pretreatment lysis method in cell lysis and BoonFix® storage experiment. Boxes extend from the lower quartiles to the upper quartiles with median values indicated by the line within each box. Whiskers represent maximum and minimum values, excluding any outliers (values indicated by circles which lie outside 1.5 times the interquartile range). Significant differences (paired t-test with Bonferroni correction) between methods are starred (* P <0.05; ** P ≤0.01; *** P ≤0.001). [Modified from Gill et al 2016 to include "BF" samples]

with lysozyme only followed by mechanical lysis (method LTL, median yield = 12 ng/µl; Figure 2.9). Since the input volume of sample used was equal in every extraction, the measured DNA concentration can be used to directly compare totalgenomic DNA yield obtained by each method. Repeated measures ANOVA showed that the difference in DNA concentration obtained using the four different lysis methods and BoonFix® storage was significant (P <0.0001). Pairwise comparisons showed that storage in BoonFix® medium at room temperature (method BF) produced a significantly higher DNA yield than aliquots extracted directly from frozen using 30 min pretreatment lysis with lysozyme only (P = 0.036 and P = 0.030, for replicate runs LN1 and LN2, respectively), aliquots extracted by extended lysis with lysozyme (LON; P = 0.024) or aliquots extracted by enzymatic lysis with lysozyme combined with bead beating (LTL; P = 0.016). Furthermore, enzymatic lysis with lysozyme combined with bead beating (LTL) produced a

significantly lower DNA yield than lysis with the enzyme cocktail (EC; P = 0.0004) or 30 min lysis with lysozyme only (P = 0.034 and P = 0.004, for replicate runs LN1 and LN2, respectively). All other comparisons were not statistically significant at a significance level of 0.05 (after Bonferroni correction).

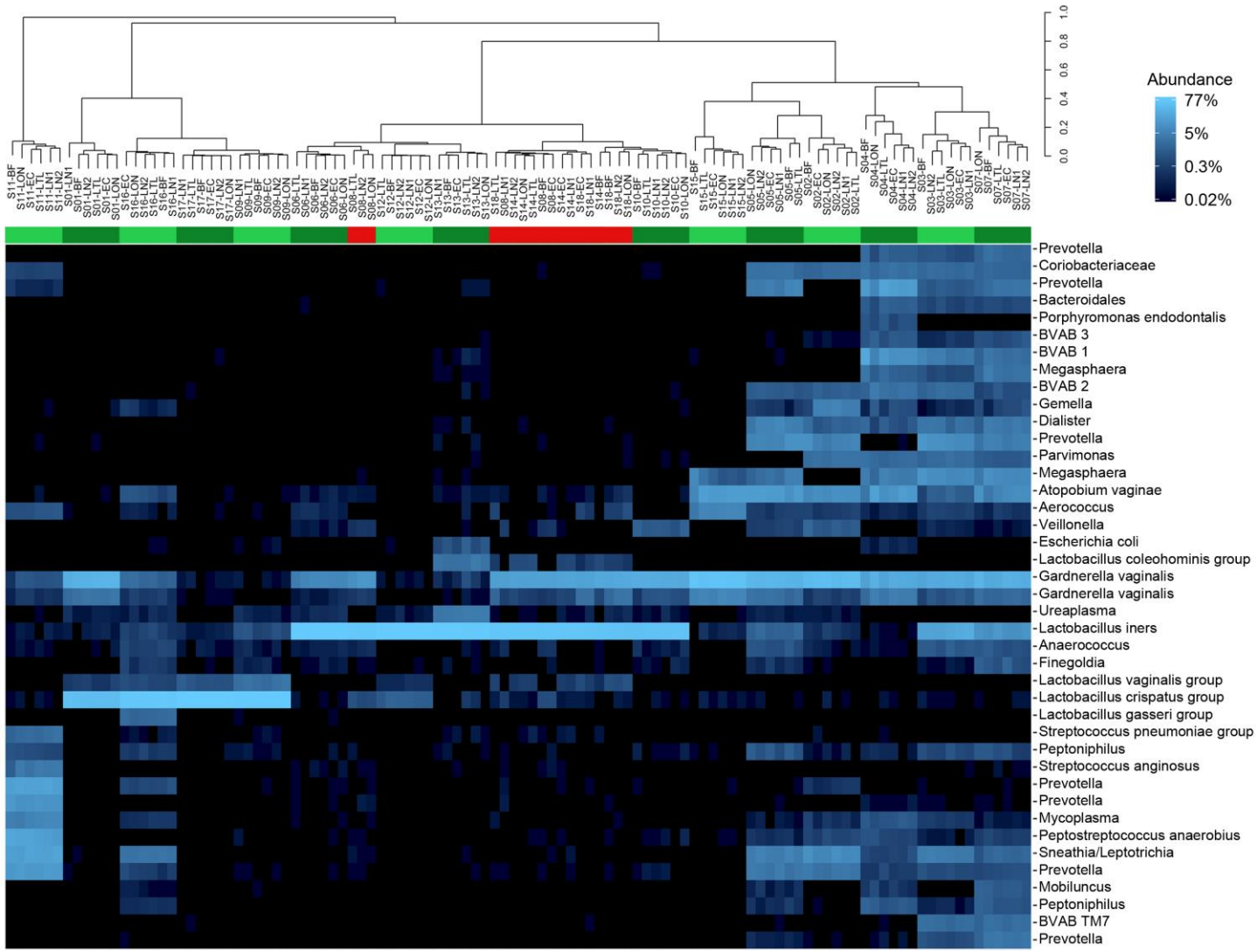
2.7.2 Vaginal bacterial community composition

The samples extracted in this study had been selected to represent a variety of microbiota profiles based on previously obtained microarray data (Borgdorff et al 2014). As expected, bacterial community profiles obtained for each extract in this study were either low in bacterial diversity (dominated by either *L. crispatus* group or *L. iners* with or without *G. vaginalis*), or high in bacterial diversity containing a mixture of strict and facultative anaerobes including *G. vaginalis*, *A. vaginae*, *Prevotella*, *Aerococcus*, *Anaerococcus*, *Streptococcus*, *Sneathia*, *Ureaplasma*, *Megasphaera*, *Mycoplasma*, *Gemella* and the bacterial vaginosis associated bacteria BVAB1, BVAB2, *Mageeibacillus indolicus* (BVAB3) and BVAB TM7 (Figure 2.10).

2.7.3 Effect on observed alpha diversity

The presence or absence of OTUs was consistent between extracts produced from the same sample using the different lysis methods and BoonFix® storage, with any discrepancies arising due to low abundance OTUs that made up no more than 0.2% of any extract, and in 93% of those cases made up less than 0.1%. There was no statistically significant difference in the number of observed OTUs between different lysis and storage methods (repeated measures ANOVA; P = 0.58). Calculation of the Simpson Index (1-D) confirmed a wide range of diversities (range = 0.02-0.89). Furthermore, the degree of variation between extracts from the same sample was small (maximum difference between extracts = 0.15). There was no statistically significant difference in the Simpson Index between the different methods (repeated measures ANOVA; P = 0.32). For the two monoculture positive controls, the Simpson index behaved as expected (0.00 in both cases), but the number of observed OTUs was distorted by singletons for one of these controls (observed OTUs = 1 and 6), illustrating that the Simpson index is a more accurate alpha diversity measure for this type of data as it is not sensitive to rare OTU read counts, which may represent read errors or incorrect sample assignment.

Figure 2.10 Heat map showing most abundant (1% or higher in at least one extract) operational taxonomic units with extracts arranged by UPGMA clustering on the Bray-Curtis dissimilarity matrix for cell lysis and BoonFix® storage experiment. Extracts are named according to the sample of origin followed by the lysis and storage method used. The coloured bar indicates which extracts have clustered most closely with all other extracts from the same sample (green) and those that have not (red). *Lactobacillus* species that could not be identified to species level at the 97% cut-off have been assigned to subgroups: *L. gasseri* group (includes *L. gasseri* and *L. johnsonii*), *L. crispatus* group (includes *L. acidophilus*, *L. helveticus*, *L. gallinarum*, *L. crispatus*, *L. jensenii* and *L. delbruekii*), *L. vaginalis* group (includes *L. vaginalis* and *L. reuteri*) and *L. coleohominis* group (includes *L. coleohominis* and *L. pontis*).



2.7.4 Effect on observed beta diversity

Between extract diversity was calculated using Bray-Curtis similarity and ranged from 57.4-99.7% within samples and from 0.0-99.9% between samples. The mean difference between replicate extractions LN1 and LN2 was 4.0% (range 0.5-11.0%). Differences between extracts from the same sample were due to differences in proportions of OTUs, rather than differences in the presence/absence of OTUs.

There was a negative correlation between the minimum within-sample Bray-Curtis similarity and the mean number of observed OTUs for that sample (Spearman's rank correlation: r = -0.67; P = 0.002). In other words, samples with higher OTU richness tended to have increased dissimilarity between extracts.

PERMANOVA analysis of Bray-Curtis dissimilarity showed that the differences between extracts originating from different samples (R2 = 0.99, P = 0.001) were far greater than differences between different lysis methods (R2 = 0.001, P = 0.002). Although the effect of lysis method was significant in this analysis, the magnitude of this effect was negligible when compared to the differences due to the sample of origin. This is reflected in the hierarchical clustering of the extracts based on Bray-Curtis dissimilarity scores (Figure 2.10) and the clustering of extracts by principal coordinate analysis ordination of the Bray-Curtis dissimilarity matrix (Figure 2.11), which resulted in clustering of the extracts by sample rather than lysis method. Pairwise comparisons revealed that the most significant differences were between methods BF (storage in BoonFix®) and methods LN2 (PERMANOVA; R2 = 0.0019, unadjusted P value = 0.004) and between method BF and LON (PERMANOVA; R2 = 0.0036, unadjusted P value = 0.009), but these differences were not statistically significant after adjustment for multiple testing (Holm-Bonferroni correction).

The largest cluster of extracts was composed of samples that were dominated by *L. iners* (66-99%) with a variable proportion of *G. vaginalis* (0-32%). In this group, several sets of extracts (from samples S08, S14 and S18) did not form discrete sub-clusters, as a result of higher Bray-Curtis similarity with extracts of other samples. This is due to small differences in observed proportions of OTUs and has occurred because of the high degree of similarity between the seven samples in this cluster. The Bray-Curtis similarity score ranged from 66-99% between extracts from different samples in this group. Since the composition of these samples was similar, we repeated the PERMANOVA analysis on this subset alone to minimise any effect

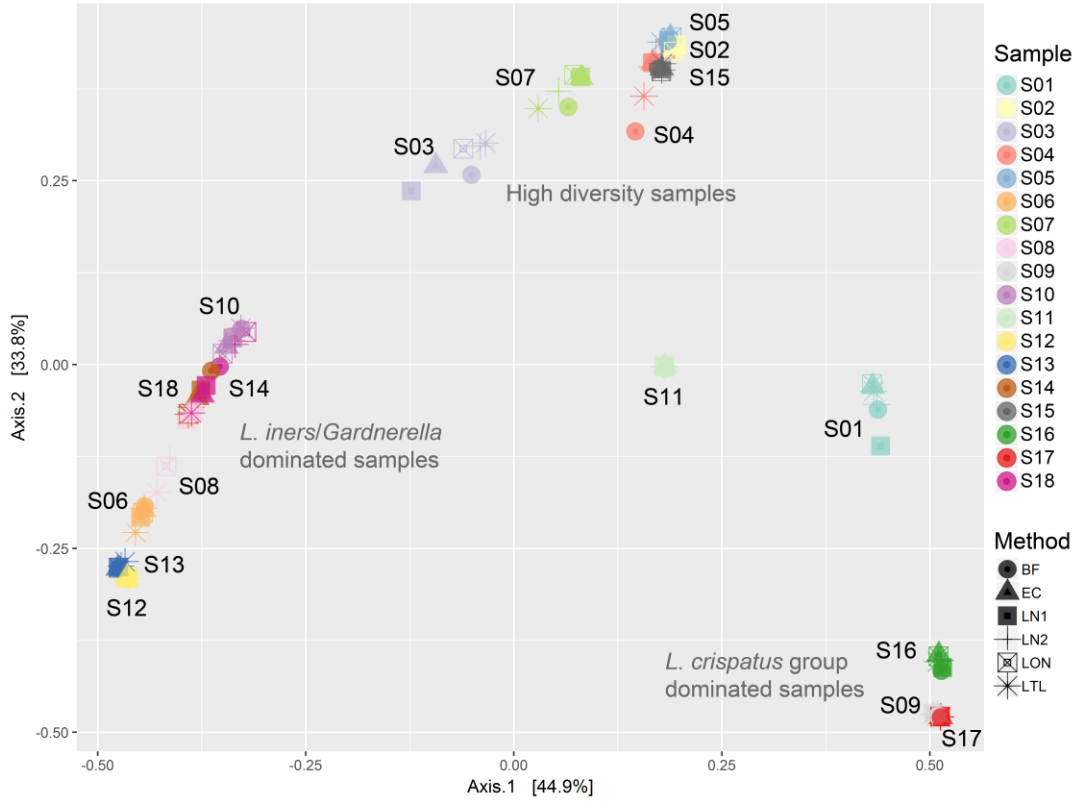


Figure 2.11 Principal coordinate analysis ordination of a Bray-Curtis dissimilarity matrix for cell lysis and BoonFix® storage experiment. Extracts are coloured by sample of origin. Extracts cluster closely with other extracts originating from the same sample and there is no observable effect of pretreatment lysis method. Extracts from samples that are dominated by *Lactobacillus iners* with variable proportions of *Gardnerella* have clustered on the left, extracts from samples that are dominated by *L. crispatus* group have clustered on the bottom right and extracts from high diversity samples that contained a mixture of strict and facultative anaerobes cluster towards the top.

of differences in alpha diversity on the magnitude of beta diversity scores. In this analysis, the differences due to sample remained highly statistically significant (R2 = 0.99, P = 0.001), but differences between different lysis and storage methods did not (R2 = 0.001, P = 0.26).

2.7.5 Effect on individual OTUs

Certain bacterial species have previously been reported to be resistant to lysozyme, including *Neisseria gonorrhoeae* (Rosenthal et al 1982) and staphylococci (Schindler and Schuhardt 1964). Furthermore, the results of a recent study indicate that streptococci may be underestimated in microbiota analyses (Brooks et al 2015). In order to investigate whether different lysis methods influenced the proportions of these bacteria, OTUs assigned to these taxa were identified and compared between different methods. Since these taxa were present at very low levels, calculations were performed on proportions calculated from raw read counts (i.e. prior to

rarefaction). One OTU identified in this study was assigned to the genus *Neisseria*. This could not be identified to species level due to 100% sequence similarity of related species in this region of the 16S rRNA gene, but is most likely to represent *N. gonorrhoeae* in this niche. This OTU was present at a low level (≤0.19%) in extracts from 5 different samples. Despite the low levels of this OTU, all of the extracts from these samples contained reads mapping to this OTU, with the exception of the sample with the lowest relative abundance (≤0.009%), where there were no reads in the extract that had been lysed with lysozyme overnight. There was no statistically significant difference between the percentage of this OTU between different lysis and storage methods (repeated measures ANOVA; $P = 0.54$). Two OTUs mapped to *Streptococcus* spp. These were consistently present in five and six samples, but were both present in only one sample at >0.1% (ranging from 1.3-2.9% and 1.0-4.1% in this sample), with all remaining sample extracts containing less than <0.08% of either OTU. There was no statistically significant difference between the percentage of this OTU between different methods for either OTU (repeated measures ANOVA; $P \geq 0.4$). *Staphylococcus* spp. were not present in the dataset after removal of small OTUs.

2.8 Discussion

Previous studies on the VMB have used either vaginal or cervical swabs (Aagaard et al 2012, Brotman et al 2012, Brotman et al 2014, Chaban et al 2014, Fettweis et al 2014, Forney et al 2010, Gajer et al 2012, Ghartey et al 2014, Hickey et al 2013, Huang et al 2015, Human Microbiome Project Consortium 2012, Jayaram et al 2014, Ling et al 2010, Martin 2012, Ravel et al 2011, Romero et al 2014, Schellenberg et al 2011, Smith et al 2012, Srinivasan et al 2012, Walther-António et al 2014), or cervicovaginal lavage samples (Frank et al 2012, Mehta et al 2015, Spear et al 2008, Spear et al 2011, Ursell et al 2014). In order to prevent post-sampling changes in the microbiota composition, most studies elected to freeze samples at -70°C or below shortly after sampling. Very few studies have utilised samples stored for any length of time in a fixative at room temperature (Chaban et al 2014, Martin et al 2012). The samples for the work presented in Chapter 4 were made available from a large epidemiological study investigating the efficacy and cost effectiveness of screening methods for cervical cancer in women with HIV (the HARP study - see Chapter 4 for a more detailed description of this project). These samples consisted of Dacron swabs stored in the coagulant fixative "BoonFix®" which contains ethanol, low molecular weight polyethylene glycol and acetic acid.

Although this medium was originally formulated as a fixative for histology and cytology, it has been used to characterise the VMB by microarray (Dols et al 2011) and detect bacterial DNA by PCR amplification (Klomp et al 2008). However, the use of alcohol-based fixatives is not widespread and little has been published about their use, in particular for molecular studies (Van Essen et al 2010). Therefore, in order to validate this storage method for microbiome analysis, we compared eight samples stored in BoonFix® at ambient temperature in South Africa to samples stored in PBS-methanol at -80°C. Although there was a tendency for the BoonFix® microbiome profiles to have slightly less Bray-Curtis similarity with the two PBS-methanol profiles, overall the results were comparable. This difference may have been due to the two samples having been taken from different sampling sites, since the BoonFix® sample was taken from the lateral vaginal wall, while the PBS-methanol sample came from the cervix. Other studies have found that these two sampling sites are very similar (Anahtar et al 2015, Huang et al 2015, Smith et al 2014, Virtanen et al 2017), but some degree of variation between these sites is not unexpected. It should be noted that the PBS-methanol samples contained a high degree of contaminants and these were most likely already present in the medium when sampling took place, since significant growth of bacteria after freezing is unlikely. These were easy to identify in this case because environmental bacteria were not expected to be present in these samples. However, it does highlight the importance of appropriate media selection for microbiome work, as such contaminants cause wasted sequencing effort at best and may lead to erroneous conclusions at worst. In order to further validate storage in BoonFix®, we compared the microbiome profiles of 18 cervicovaginal lavage samples before and after a 7.5 month period of storage in BoonFix® at room temperature. This showed that, although there were some differences in beta-diversity, these differences were minor compared to the differences between women. Interestingly, with these samples, storage in BoonFix® increased DNA yield. The reason for this is not known. However, since ethanol can decrease bacterial membrane integrity (Da Silveira et al 2002), this could relate to improved cell lysis. If this were to be the case, there is potential for this fixative to improve DNA extraction from difficult-to-lyse bacteria and thereby improve the accuracy of microbiome profiling, but this theory would require further testing. Overall, these results show that cervicovaginal samples stored in BoonFix® are suitable for microbiome analysis and the results are comparable to samples stored frozen. Furthermore, significant sample contaminants were not detected in BoonFix® medium in either study, which is consistent with it having bacteriocidal (or bacteriostatic) properties and hence

preventing distortion of the microbiome after sampling in samples stored at room temperature.

One potential problem encountered using the samples stored in BoonFix® was that the DNA yield was initially very low (see part I). Recent work has highlighted that, in studies using next-generation sequencing technology to characterise the bacterial microbiota, low amounts of template DNA are associated with a proportional increase in contaminant taxa originating from laboratory reagents (Biesbroek et al 2012, Kennedy et al 2014, Salter et al 2014) and this is consistent with our own results (see Appendix B). These contaminants can lead to erroneous conclusions. Furthermore, we showed that amplicon yield is dependent on DNA extract concentration and good amplicon yield is necessary to allow accurate equimolar pooling of amplicons prior to sequencing. It was therefore important to optimise DNA extraction from BoonFix® samples and we tested whether inclusion of the vaginal swab in the extraction process could improve DNA concentration of extracts. We found that DNA yield could be significantly increased by including the swab head in the proteinase K/"buffer AL" digestion step, which was then at a level considered sufficient to avoid significant reagent contamination for most samples (see Appendix B). This result is consistent with experiments on human DNA yield from forensic swabs, for which the inclusion of the swab head in the proteinase K/"buffer AL" digestion step is routine when using the QIAamp DNA Investigator kit, also made by Qiagen (Adamowicz et al 2014). A study by Adamowicz and others (2014) found that DNA yield from swab samples can further be improved by increasing the length of the proteinase K digestion step (to between 3 and 18 hours) and by performing periodic resuspension of the swab (centrifuging the swab in a spin basket and collecting the eluate in the same sample tube used in the extraction process) (Adamowicz et al 2014). However, yield was deemed sufficient for downstream analyses and further optimisation was therefore not necessary in this study. In part III of this chapter, 10 ng of template DNA was used in each 25 µl PCR reaction. This amount of DNA has been found to result in significantly lower variability in microbiota community structure in studies profiling the faecal microbiota (Kennedy et al 2014). Furthermore, in our laboratory, this concentration of DNA in cervicovaginal samples resulted in negligible levels of reagent contamination (see Appendix B), which is supported by the low levels of the contaminant *Rhodanobacter* OTU in this study. *Rhodanobacter* spp. have been isolated from environmental soil and water samples (Hemme et al 2015, Van Den Heuvel et al

2010). Interestingly, this genus has also been reported as a member of the human microbiota from studies using Qiagen extraction kits (Audirac-Chalifour et al 2016, Weng et al 2014), highlighting the need to adequately control for contamination occurring during laboratory processing of samples for 16S rRNA microbiota profiling.

A further variable that may distort microbiome profiles is differential lysis efficiency of the various species that may be present in vaginal samples. Previous studies have used mock communities to be able to assess different lysis methods (Abusleme et al 2014, Yuan et al 2012). Mock community studies have the advantage that the community profile of the samples is known, allowing assessment of the accuracy of the results (Yuan et al 2012). In this study, we chose to use naturally occurring bacterial communities for which the true composition was not known and as a result the accuracy of each lysis method could not be determined. However, using biological samples that cover a range of different community types has the advantage of allowing comparison of lysis and storage methods on a wider range of bacteria commonly encountered in the vaginal niche, including those that have not or have rarely been cultured. This includes the bacterial vaginosis-associated bacteria, which can make up a substantial proportion of the bacterial population in some individuals (Oakley et al 2008). Additionally, using vaginal samples allowed us to compare the magnitude of the effect of method with that resulting from biological differences between samples from different individuals. It should also be noted that vaginal samples can vary in consistency (Chappell et al 2014) and may contain viscous mucoid material that is difficult to homogenise. In this study, we have chosen to remove any such material where present prior to processing to minimise any potential variation between extracts resulting from inadequate homogenisation. It is possible that the composition of the microbiota associated with the removed material differed from the remaining material and could therefore have changed the overall profile of the samples. However, as with the effect of storage in BoonFix®, the results of this study clearly show that sample has a far greater effect on the microbiota profile than the pretreatment lysis method. This is consistent with the results of studies that have compared different extraction kits or protocols for faecal samples (Salonen et al 2010, Wagner Mackenzie et al 2015, Wesolowska-Andersen et al 2014) and saliva (Willner et al 2012). Additionally, epidemiological studies investigating the effect of vaginal bacterial communities on health commonly group samples by clustering based on overall community

structure, assigning each sample to a “community type” (Borgdorff et al 2014, Brotman et al 2014, Frank et al 2012, Hummelen et al 2010, Lee et al 2013, Ravel et al 2011) and accurate clustering of extracts was not affected by either pretreatment lysis method or BoonFix® storage in this study. However, biological differences resulting from subtle variation in proportions of taxa may be difficult to separate from experimental variation as evidenced by up to 11.0% dissimilarity between replicate extracts LN1 and LN2, and should therefore be interpreted with caution. A larger sample size and greater number of experimental replicates would be required to investigate this variation further, particularly to determine whether the other lysis methods and BoonFix® storage used in this study would produce a similar degree of dissimilarity.

In comparing pretreatment lysis methods, we used the recommended protocol for the pretreatment of Gram-positive bacteria with lysozyme as standard since it is thought to improve species representation (Abusleme et al 2014). As a result, we cannot make any conclusions about the necessity of this enzyme for bacterial lysis from the data in this study. Lysozyme breaks down the bacterial cell wall by cleaving peptidoglycan and may be particularly important for the breakdown of the thick peptidoglycan layer of the Gram-positive cell wall (Davis and Weiser 2011). However, modifications to peptidoglycan structure can render bacteria resistant to lysozyme digestion. This has been reported in *N. gonorrhoeae* (Rosenthal et al 1982) and staphylococci (Schindler and Schuhardt 1964), both of which could be present in vaginal samples (Donders et al 2002, Wiesenfeld et al 2003). The use of the enzymes mutanolysin and lysostaphin in addition to lysozyme has been recommended for the lysis of vaginal samples in order to lyse bacterial species that are resistant to lysozyme digestion (Yuan et al 2012). Lysostaphin specifically lyses some *Staphylococcus* spp. (Schindler and Schuhardt 1964) and mutanolysin is active against the cell wall of some streptococci (Yokogawa et al 1974). In this study, we failed to identify any differences between lysis methods for the aforementioned bacterial taxa. However, the number of 16S rRNA reads mapping to these genera was small, resulting in low statistical power to detect relatively small differences. Additionally, the bacterial species/strains sequenced in this study may not have been resistant to lysozyme lysis. For example, differences in susceptibility to lysozyme digestion between different strains of *N. gonorrhoeae* have been reported (Rosenthal et al 1982). It is possible that differences in lysis efficiency may have been evident if different species or strains had been present in the samples

used. However, the addition of mutanolysin and/or lysostaphin to samples in this study, which contained the majority of major vaginal bacterial taxa, was not found to significantly alter the presence/absence of OTUs or their relative abundance. It is therefore unlikely that the addition of these enzymes would alter the conclusions of studies designed to investigate the impact of vaginal community type on human health. A further additional treatment that can be used to improve lysis of cells is mechanical disruption, usually by bead-beating. Bead beating has been reported to increase the observed richness in previous microbiota studies (Guo and Zhang 2013, Salonen et al 2010). This was not the case in this study in which we found no significant difference in alpha diversity. It should be noted that bead-beating may have a greater influence on fresh samples compared with those that have been stored in the freezer, possibly due to disruption of the Gram-positive cell wall by freeze-thawing (Wesolowska-Andersen et al 2014). The samples used in this study were stored at -80°C, as is common for vaginal microbiota studies (Brotman et al 2012, Brotman et al 2014, Forney et al 2010, Frank et al 2012, Huang et al 2015, Jayaram et al 2014, Ling et al 2010, Mehta et al 2015, Ravel et al 2011, Schellenberg et al 2011, Smith et al 2012, Walther-António et al 2014) and it is possible that an effect of bead beating would have been evident if fresh samples had been used, by resulting in reduced richness in those extracts that were not subjected to bead beating.

In contrast to the effect on diversity, we found that the addition of a bead-beating step significantly reduced the DNA concentration of the extract, which is consistent with previous results using mock bacterial communities (Abusleme et al 2014, Yuan et al 2012), and is most likely due to some material being lost with the beads when they are removed from the sample. DNA yield is commonly used to assess the efficiency of different lysis and extraction protocols. Other studies have reported that the inclusion of a bead-beating step led to an increase in DNA yield from activated sludge (Guo and Zhang 2013) and faecal samples (Maukonen et al 2012). However, these samples may be more heterogeneous and particulate in nature, which could explain this difference and emphasises the importance of validating methods for microbiota analysis on the sample type of interest. Consistent with our results, higher DNA yield has not been shown to be associated with improved accuracy of microbiota profiles in mock community studies (Abusleme et al 2014, Yuan et al 2012), and the reduced DNA yield with method LTL is not of particular concern. However, increased contamination caused by inclusion of a bead-beating

step has been reported (Abusleme et al 2014) and may be best avoided in the absence of a clear advantage. It should be noted that DNA extraction to test the effect of different pretreatment lysis strategies was carried out with the same type of commercial kit (Qiagen DNeasy Blood and Tissue) for all samples. Proprietary extraction kits employ a variety of different techniques to lyse cells and purify DNA. Hence the importance of pretreatment with additional lysis methods may vary between kits.

2.9 Conclusions

It is widely acknowledged that bias exists in 16S rRNA studies describing microbiota profiles and that no currently available method is able to perfectly describe the community being analysed. However, an understanding of how the choice of laboratory methods affects the results of such studies is important in order to accurately interpret the results and make valid comparisons between different studies. Although we were able to identify statistically significant differences in DNA yield and diversity between the different lysis methods and BoonFix® storage, the effects of this were much smaller than those due to the individual variation between subjects, and did not alter the grouping of extracts by hierarchical clustering and principal coordinate analysis.

Going forward, it was decided to use the Qiagen Blood and Tissue Kit for the extraction of BoonFix® samples with inclusion of the swab in the proteinase K step, since this produced the best DNA yields and was therefore less likely to be associated with significant reagent contamination of microbiome profiles. The recommended pretreatment for Gram-positive bacteria as per the kit's handbook (treatment with enzymatic lysis buffer containing lysozyme) was decided upon as the results were comparable to the other, more involved, lysis methods tested.

CHAPTER 3: Bioinformatics Methods

3.1 Introduction

Studies using next generation sequencing of 16S rRNA gene amplicons generate several million DNA sequence reads. While this allows for deep sampling of the microbiota, it also produces a large amount of data that needs to be summarised *in silico* (i.e. performed on a computer) before diversity can be calculated and statistical tests can be performed to determine whether there is a significant association between the microbiota profiles and health outcomes. In order to condense data into a manageable form, reads are grouped based on their DNA sequences. One of the challenges with this is that 16S rRNA sequencing data contains errors. In Illumina data, sequence error is created through base substitutions during PCR, the formation of PCR chimeras and incorrect base calling by the sequencer (Tikhonov et al 2015). PCR substitution error can be reduced through the use of a proofreading DNA polymerase which has the ability to excise incorrectly paired bases (Gohl et al 2016) and can be partially mitigated *in silico* by performing an error correction step (Schirmer et al 2015). Chimeric sequences are created during PCR from two different templates and this type of error can be reduced by restricting the number of PCR cycles, usually to 25 or less (Kanagawa 2003) and is also affected by the choice of DNA polymerase (Gohl et al 2016). Any chimeric sequences that do form should be removed after sequencing *in silico*. Sequencing error can only be dealt with *in silico* and is perhaps the biggest challenge when grouping or "clustering" reads. The most common type of sequencing error in Illumina data are base substitutions (Schirmer et al 2015).

The first algorithms used to condense 16S rRNA sequencing data cluster reads into "operational taxonomic units" (OTUs) – a process also referred to as "OTU picking" – include USEARCH and CD-HIT. These work by randomly selecting a read as the OTU centroid, and then consigning all similar sequences to that OTU, based on an arbitrary global similarity threshold to the selected centroid, commonly 97% (Edgar 2010, Fu et al 2012). In this context "global" refers to the similarity between two DNA sequences along their entire length. The centroid sequence may be picked either by the algorithm (termed *de novo* OTU picking since they do not require user input) or selected from a user-defined database (closed OTU picking), or a combination of the two (open OTU picking). These approaches are “greedy” because they use a relatively simple strategy to solve a more complex problem by utilising a local optimum (i.e. picking a centroid) in the hope of finding the global

optimum (i.e. that this centroid will produce an accurately delineated OTU). For these algorithms, employing a similarity threshold is necessary due to the presence of error in the sequencing results which artificially decreases pairwise sequence similarity and would thereby – without the use of a similarity threshold – erroneously inflate diversity estimates. However, while this approach is simple to implement computationally, it has some disadvantages. For example, some experts argue that it does not accurately reflect biological diversity (Eren et al 2016). Organisms evolve at different rates such that slowly evolving lineages share a high degree of pairwise sequence similarity while those that evolve more quickly have a lower degree of similarity. Hence, by using the same threshold for all lineages, bacterial species that have evolved slowly are separated poorly, while those that have evolved more quickly are insufficiently clustered (Koeppel and Wu 2013). In the vaginal niche, differentiation of the *Lactobacillus* spp. is of particular interest and many species cannot be resolved using a 97% similarity threshold (and in many cases not even with a 98 or 99% similarity threshold, see Appendix E). Furthermore, by picking an arbitrary centroid it is possible that closely related sequences are placed in different OTUs (Mahé et al 2014).

More recently, new *de novo* algorithms have been designed that aim to address some of these issues and better describe biological variation, despite the presence of read error. This increased level of detail has the potential to provide more information about population structure that may have clinical and epidemiological relevance (Eren et al 2011). These methods include oligotyping (Eren et al 2013), Swarm (Mahé et al 2014) and DADA2 (Callahan et al 2016). Oligotyping works by using Shannon entropy (a measure of information uncertainty) to identify positions within the sequence data where there is true biological variability (i.e. where the variation is greater than would be expected due to sequencing error) and uses this information to bin sequences into groups. To differentiate these groups from those produced using methods that rely on a global similarity threshold the authors use the term "oligotypes" in place of OTUs (Eren et al 2013). Swarm uses information from pairwise sequence alignment, but rather than comparing to a centroid, comparisons are made to all other reads in the dataset and sequences are included in the same cluster if they are all linked to one another by sequences that differ by no more than d bases (where a difference is defined as either a substitution, insertion or deletion; d is set to 1 by default). Clusters generated in this way are then further refined by identifying weak links within a cluster, i.e. where the number of

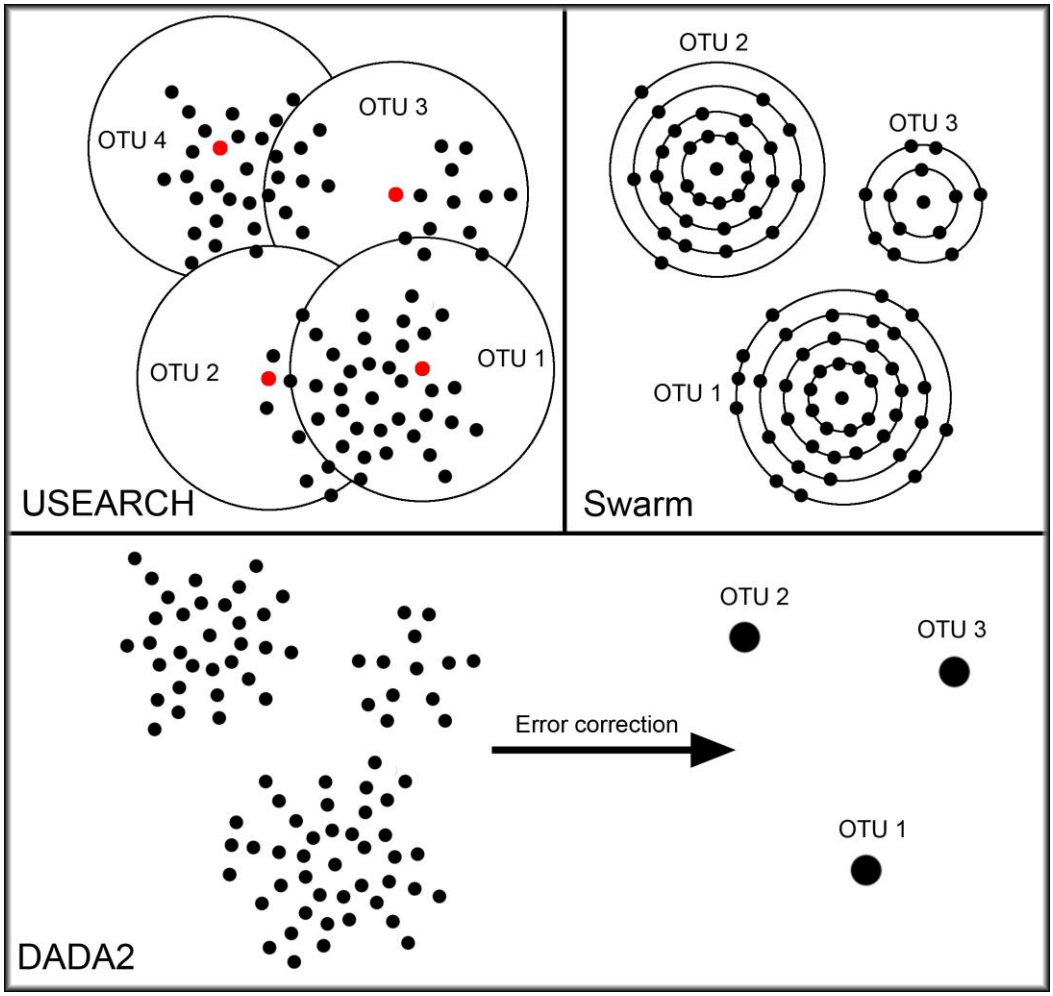


Figure 3.1 Schematic summary of clustering strategies employed by USEARCH, Swarm and DADA2. Individual reads are indicated by dots as they might appear on a PCoA plot. USEARCH selects an (arbitrary) centroid, indicated in red. All reads that are within a given similarity threshold (t; often set to 97%) from that centroid are assigned to the same OTU. The diagram illustrates the potential problem with this approach where the dataset contains multiple closely related sequence clusters. Swarm uses a local clustering threshold (d; usually set to 1) and iteratively adds reads to an OTU if they differ by no more than d bases. DADA2 attempts to correct read error prior to assigning reads to OTUs, and reads are assigned to the same OTU is their corrected sequences are identical. Adapted from [Mahe 2014].

sequences that link two parts of a cluster is low, and separating these parts in a process termed "OTU breaking" (Mahé et al 2014). DADA2 infers expected error rates from sequence composition and quality data, defining the rate at which erroneous sequences are expected to be produced from each parent sequence. This information is then used to create clusters that contain the parent sequence as well as the expected number of associated erroneous sequences determined by the error model (Callahan et al 2016). To decide which method of sequence clustering was most appropriate for analysis of the vaginal microbiome (VMB) in the study described in the following chapters, we tested Swarm v. 2.1.13 and DADA2 v. 1.4.0

on a control dataset and compared the results to those obtained by USEARCH v. 6.1.544, with and without a reference database (Figure 3.1).

3.2 Methods

In order to test the accuracy and efficiency of clustering methods for 16S rRNA data, we used a set of controls as follows.

3.2.1 Description of controls

Three types of bacterial DNA sources were utilised in this study: the commercially available Zymo Microbial DNA Standard (Zymo Research, Irvine, USA), DNA extracts of cultured vaginal bacteria and clinical vaginal samples. In addition, nuclease-free water was extracted using the same DNA extraction kit for use as a negative control to identify reagent contaminants.

The Zymo Microbial DNA Standard contains pooled DNA extracted from pure cultures of eight bacterial species (*Pseudomonas aeruginosa*, *Escherichia coli*, *Salmonella enterica*, *Lactobacillus fermentum*, *Enterococcus faecalis*, *Staphylococcus aureus*, *Listeria monocytogenes* and *Bacillus subtilis*). These bacterial strains are well characterised and information on genome size, average GC content and 16S rRNA copy number is available. Each batch is created to provide a standardised theoretical composition and the results of shotgun metagenomics sequencing for each batch can be accessed, providing a gold standard measure of sample composition.

Six vaginal bacterial cultures (*Lactobacillus crispatus*, *Lactobacillus iners*, *Lactobacillus jensenii*, *Atopobium vaginae*, *Prevotella bivia* and *Gardnerella vaginalis*) were obtained from the HIV/STI Reference Laboratory, Institute of Tropical Medicine, Antwerp. These were stored in BoonFix® at room temperature, pelleted by centrifugation at 5000 x g for 10 minutes and extracted using the Qiagen DNeasy Blood and Tissue kit as previously described (see section 2.2.2). Additionally, a pure culture of *Lactobacillus amylovorus* (originally isolated from porcine intestine) was obtained from a colleague at the Institute of Integrative Biology, University of Liverpool, UK. This lactobacillus was included based on its high degree of sequence similarity to *L. crispatus* in the V3-V4 region (99.5%,

differing by two base substitutions). The sample was stored frozen at -20°C for a brief period prior to extraction using the same protocol.

Two cervicovaginal lavage sample extracts (S11 and S14) produced previously (described in section 2.6.1 and extracted according to method "LN1", see sections 2.6.2 and 2.6.3) were also sequenced.

3.2.2 Amplicon library preparation and DNA sequencing

The V3-V4 region of the 16S rRNA gene was amplified in a 25 µl reaction comprising 12.5 µl of NEBNext® High-Fidelity 2x PCR Master Mix (containing DNA polymerase Q5), 1.25 µl each of a 10 µM concentration of the conserved bacterial 16S rRNA primers 319F 5'-ACTCCTACGGGAGGCAGCAG-3' and 806R 5'-GGACTACHVGGGTWTCTAAT-3' (Fadrosh et al 2014) adapted with linker regions to allow barcoding of sequences using a dual-indexing approach (D'Amore et al 2016). The V3-V4 region was chosen since – despite having slightly reduced ability to discriminate between *Lactobacillus* spp. – this region has substantially lower levels of *Lactobacillus* intra-species sequence diversity (see Appendix E), thereby reducing the likelihood of sequences from the same species being separated and, within the lactobacilli, there is also a comparatively low degree of sequence variability within the same genome (see Appendix E). Furthermore, two other regions commonly used in VMB studies (V1-V2 and V6) tend to overestimate alpha diversity (Youssef et al 2009). PCR products were produced from 2 ng of genomic DNA for the Zymo Microbial DNA Standard (2 replicates each on 11 PCR runs). Vaginal bacteria (two replicates each) and *L. amylovorus* (single amplification) were PCR amplified separately in reactions containing 40 ng of genomic DNA for vaginal bacteria, with the exception of *L. iners* and *L. jensenii* extracts for which total DNA yield was too low (6.4 ng and 2.3 ng used, respectively). Additionally, a pool was created containing approximately equal genome copies from each of the six vaginal species (the "VMB mock community"). Genome copy number was estimated based on genome size according to the NCBI database (Acland et al 2014) and 40 ng of pooled DNA was used per PCR reaction (2 replicates each on 12 PCR runs). For the cervicovaginal lavage samples 30 ng per reaction was used (a total of 23 and 24 replicates for S11 and S14 respectively, across 12 PCR runs). PCR products from negative extraction controls were produced from 10 µl of undiluted extract (39 replicates).

For the first PCR (16S rRNA gene amplification) the samples were initially denatured at 98°C for 30 s, followed by 10 cycles of 98°C for 10 s, 58°C for 30 s and 72°C for 30 s, with a final extension at 72°C for 5 min. The PCR products were then purified using SeraPure magnetic beads (Faircloth and Glenn), before undergoing a second PCR to attach sample-specific barcodes and further amplify the region of interest. The second PCR consisted of a 25 µl reaction containing 7.5 µl of clean PCR product, 12.5 µl of NEBNext® High-Fidelity 2X PCR Master Mix and 2.5 µl each of Nextera XT Index Kit v2 (Illumina, San Diego, USA) indexing primers. Samples were initially denatured at 98°C for 3 min, followed by 15 cycles of 98°C for 30 s, 55°C for 30 s and 72°C for 30 s, with a final extension at 72°C for 5 min. PCR products were purified, eluted in a volume of 10 µl TE buffer (Sigma-Aldrich) and quantified using the Qubit Fluorometer with the dsDNA HS Assay kit to determine amplicon yield. Purified PCR amplicons measuring ≥ 2 ng/µl were run on a 2% agarose gel at 100V to verify purity of the amplicon. Additionally, to test for PCR bias, the single vaginal species were individually PCR amplified as above using four different barcode combinations and then pooled in equal concentrations prior to sequencing ("VMB mock community pooled post-PCR"; 3 replicates of each of the 4 barcode combinations sequenced). Amplicons were then pooled and sequenced at the University of Liverpool Centre for Genomics Research on the Illumina HiSeq platform (2x300bp; Illumina) on two separate runs, one consisting of two lanes on the same flowcell (designated "X" and "Y") and one consisting of a single lane (designated "Z").

3.2.3 Bioinformatics

Sequencing reads were demultiplexed and primer sequences were trimmed using Cutadapt v. 1.2.1 (Martin 2011). The resulting reads were then processed according to one of three different bioinformatics pipelines as follows:

USEARCH: Primer trimmed reads were error corrected using SPAdes v. 3.1.0 (Bankevich et al 2012) and paired-end alignment was performed using PEAR v. 0.9.6 (Zhang et al 2014) with a size cut-off of 380-480 bases. After removal of chimeric sequences detected by USEARCH *de novo* and against the Silva v. 128 database (sequences flagged by both methods removed), sequences were binned into OTUs either *de novo* or closed (i.e. against a vaginal reference database) using USEARCH v. 6.1.544 (Edgar 2010) through Quantitative Insights Into Microbial Ecology (QIIME) v. 1.8.0 (Caporaso et al 2010) with the similarity threshold set at

97% or 99%. Taxonomic assignment of representative sequences (most abundant) was carried out for each OTU by RDP classifier with a minimum bootstrap value of 50% for both closed and *de novo* clustered data (Wang et al 2007) against the Silva v. 128 database (non-redundant, clustered at 97% for OTUs created using a 97% similarity threshold and at 99% only for OTUs created using a 99% similarity threshold) (Pruesse et al 2007) in QIIME. For the OTUs created with a 97% similarity threshold, taxonomy assignment was also performed using the 99% database, to see if this could be used to improve taxonomical assignment, based on the assumption that the most abundant sequence in the OTU is most likely to be the "real" sequence. Unless otherwise stated, the reported taxonomy is the one assigned by the 97% clustered database. After clustering *de novo*, OTUs that contained less than 0.005% of total reads were removed as likely sequencing errors (Bokulich et al 2013). For closed OTU picking, a non-redundant V3-V4 reference database was created using sequence identifiers from the Vaginal 16S rDNA Reference Database (Fettweis et al 2012) which contains sequences for the V1-V3 region of vaginal bacteria. Sequences for the bacteria in the Zymo control were also included if not already in the database (i.e. *Salmonella enterica*, *Listeria monocytogenes* and *Bacillus subtilis*), as was the sequence for the *Rhodanobacter* sp. found previously in negative extraction controls.

Swarm: Primer trimmed reads were error corrected using SPAdes v. 3.1.0 (Bankevich et al 2012) and paired-end alignment was performed using PEAR v. 0.9.6 (Zhang et al 2014) with a size cut-off of 380-480 bases. The obtained sequences then underwent removal of ambiguous bases and dereplication using Vsearch v. 2.4.3 (Rognes et al 2016). Sequences were clustered into OTUs *de novo* using Swarm v. 2.1.13 (Mahé et al 2015) with the difference parameter (d) set at 1 and option "fastidious" enabled. The resulting representative sequences (most abundant) were abundance sorted and checked for chimeric sequences with Vsearch, using the UCHIME algorithm (Edgar et al 2011) *de novo* and against the Silva v. 128 database (only sequences detected by both methods removed). Taxonomic assignment of representative sequences (most abundant) was carried out for each OTU by RDP classifier with a minimum bootstrap value of 50% (Wang et al 2007) against the Silva v. 128 database (non-redundant, clustered at 99%) (Pruesse et al 2007) in QIIME. OTUs with a total read count below 100 were discarded prior to further analysis.

DADA2: Paired end sequences were trimmed to remove the poor quality ends and filtered to remove poor quality reads through the DADA2 v. 1.4.0 (Callahan et al 2016) pipeline for large paired end datasets. The length to which the forward and reverse reads were trimmed was chosen based on quality plots and later adjusted (shortened) to optimise retention of reads. At the same time the maxEE parameter (which sets the maximum number of allowed "expected errors" per read) was increased and the truncQ value set to 0, to achieve ~90% read retention. Final parameters for the forward and reverse reads were trimming to 250 bp and 230 bp with a maxEE value of 5 and 8, respectively. The error rates of forward and reverse reads were determined separately for each set of sequences from the three Illumina HiSeq lanes, as recommended by the developers. The DADA2 pipeline then proceeded with sample inference (analogous to OTU picking) and paired end alignment. Finally, obtained sequence tables were merged, *de novo* chimera checking was performed and taxonomy was assigned using RDP classifier with a minimum bootstrap value of 50% (Wang et al 2007) against the Silva v. 128 database (non-redundant, clustered at 99%) (Pruesse et al 2007). Species assignments were made using DADA2's *addSpecies* method which identifies exact sequence matches to the database. The settings were such that only reference sequences with unique matches were assigned a species classification.

3.2.4 Data analysis

The NCBI BLAST search tool (Altschul et al 1990) and Clustal Omega (Sievers et al 2011) were used to match the obtained reference sequences to those held in the NCBI database and to compare sequence similarity, respectively.

Calculation of alpha and beta diversity measures, statistical analyses and graphing of data were performed in R version 3.2.2 (R Core Team 2015) and using the vegan package version 2.3-2 (Oksanen et al 2015) and nlme package version 3.1-131 (Pinheiro et al 2017). For the Zymo Microbial DNA standard control profiles, a mixed effects linear regression model was used to identify differences in accuracy (as determined by Bray Curtis similarity of each replicate to the expected profile) between different clustering methods, with replicate ID as a random effect (to control for differences due to PCR and sequencing). Bray-Curtis similarity on relative abundance data was used to report and assess differences in beta diversity. Permutational multivariate ANOVA (PERMANOVA) (Anderson 2001) was used to

assess differences in beta diversity (Bray Curtis similarity) between different sequencing runs/lanes.

3.3 Results

A total of 17,835,288 16S rRNA sequence reads were obtained from the 106 control samples. One of the VMB mock samples failed to sequence, despite producing an adequate amount of amplicon, most likely due to a pooling mistake (i.e. the sample was not added to the final pool of amplicons that were put on the Illumina HiSeq). The raw read count per sample ranged from 43,095 to 1,856,798 reads, with a median count of 142,979 reads. Following processing through the bioinformatics pipelines, the median read count per sample in the final OTU table was highest for USEARCH closed at 97% similarity (136,377 reads), followed by Swarm (median = 132,820), DADA2 (131,812 reads), USEARCH closed at 99% similarity (128,359 reads), USEARCH *de novo* at 97% similarity (126,317 reads) and lowest for USEARCH *de novo* at 99% similarity (median = 124,288).

3.3.1 Analysis time and ease of use

Good documentation is available for the use of USEARCH within QIIME and for DADA2, providing workflows that are relatively simple to follow. Although Swarm has been incorporated into various bioinformatics pipelines such as QIIME, the newest version was only available as standalone software. This meant that assembling an OTU table from the clustered data was less straightforward than with the other clustering methods. However, an example workflow is available online (Mahé 2016).

When running a clustering algorithm on multiple core platforms, speed in real time depends not only on the platform used and how the clustering strategy is implemented, but also on whether it allows multithreading (i.e. whether the code can be run on several processors simultaneously). USEARCH is reported to be the fastest greedy clustering algorithm available and can be faster than Swarm (Mahé et al 2014). However, when clustering *de novo*, USEARCH is not capable of multi-threading, meaning it can only use one processor/thread at a time. Therefore, in real time, Swarm is often still faster than USEARCH, depending on the number of processors/threads available. Additionally, when clustering *de novo* with a 99% similarity threshold, USEARCH used up more CPU time compared with Swarm in this study (Table 3.1). In contrast, when clustering against a database, USEARCH

does allow multi-threading and is very fast. DADA2 is comparable to Swarm and USEARCH in terms of CPU time taken up and is also capable of multithreading, reducing real processing time.

It should also be mentioned that *de novo* chimera checking took a considerable amount of time in the Swarm (48 hours (h) 31 minutes (m) 29 seconds (s) using Vsearch) and USEARCH (79h54m56s using USEARCH) pipelines. In these pipelines the number of reads that were checked was relatively high. By comparison, the DADA2 pipeline is considerably faster (8m1s in real time). This is probably due to the fact that DADA2 performs this check only on the relatively small set of representative sequences remaining after filtering of error prone reads and error correction (see 3.2.3).

Table 3.1 Time taken for each OTU clustering algorithm to run. Real time refers to actual time elapsed, user and system time are the actual CPU time taken up with executing the process in user mode and within the kernel, respectively. Note that DADA2 processing also includes paired end alignment, which was performed with PEAR prior to Swarm and USEARCH.

ALGORITHM	REAL TIME	USER TIME	SYSTEM TIME
DADA2: Calculating error rates	82m9s	225m45s	4m30s
DADA2: Sample inference, paired end merging	35m34s	105m6s	1m30s
Swarm: Clustering	94m59s	460m12s	11m39s
USEARCH denovo 97%: Clustering and generation of OTU map	350m33s	333m55s	0m56s
USEARCH denovo 99%: Clustering and generation of OTU map	1044m35s	1029m30s	1m45s
USEARCH closed 97%: Clustering and generation of OTU map	12m1s	0m42s	0m35s
USEARCH closed 99%: Clustering and generation of OTU map	11m4s	6m51s	0m33s

3.3.2 Zymo Microbial DNA Standard: Species identification and accuracy

For the Zymo Microbial DNA Standard, the overall profiles obtained by each method were very similar (Figure 3.2). The representative sequences for the eight most abundant sequence clusters identified by each method were identical to the expected DNA sequence (as provided by Zymo). However, although this usually resulted in correct taxonomical assignments by RDP classifier, in some cases, *Bacillus subtilis* was identified as *Bacillus mojavensis* (see Figure 3.2). This has occurred because both species have the same DNA sequence in this region and illustrates the stochastic nature of this classification process. Since DADA2 uses exact matches for species identification, this has not occurred with this method.

Both DADA2 and Swarm generated small but sizeable additional sequence clusters: both produced a sequence cluster consisting of *Salmonella* sequences (1.4-1.6% in both methods), with the reference sequence differing from the expected sequence by two base substitutions. In addition, DADA2 generated a *Bacillus* sequence cluster (2.0-2.5%) with the reference sequence differing from the expected sequence by a single base substitution. Swarm is not expected to be able to separate sequences differing by a single nucleotide which explains why Swarm did

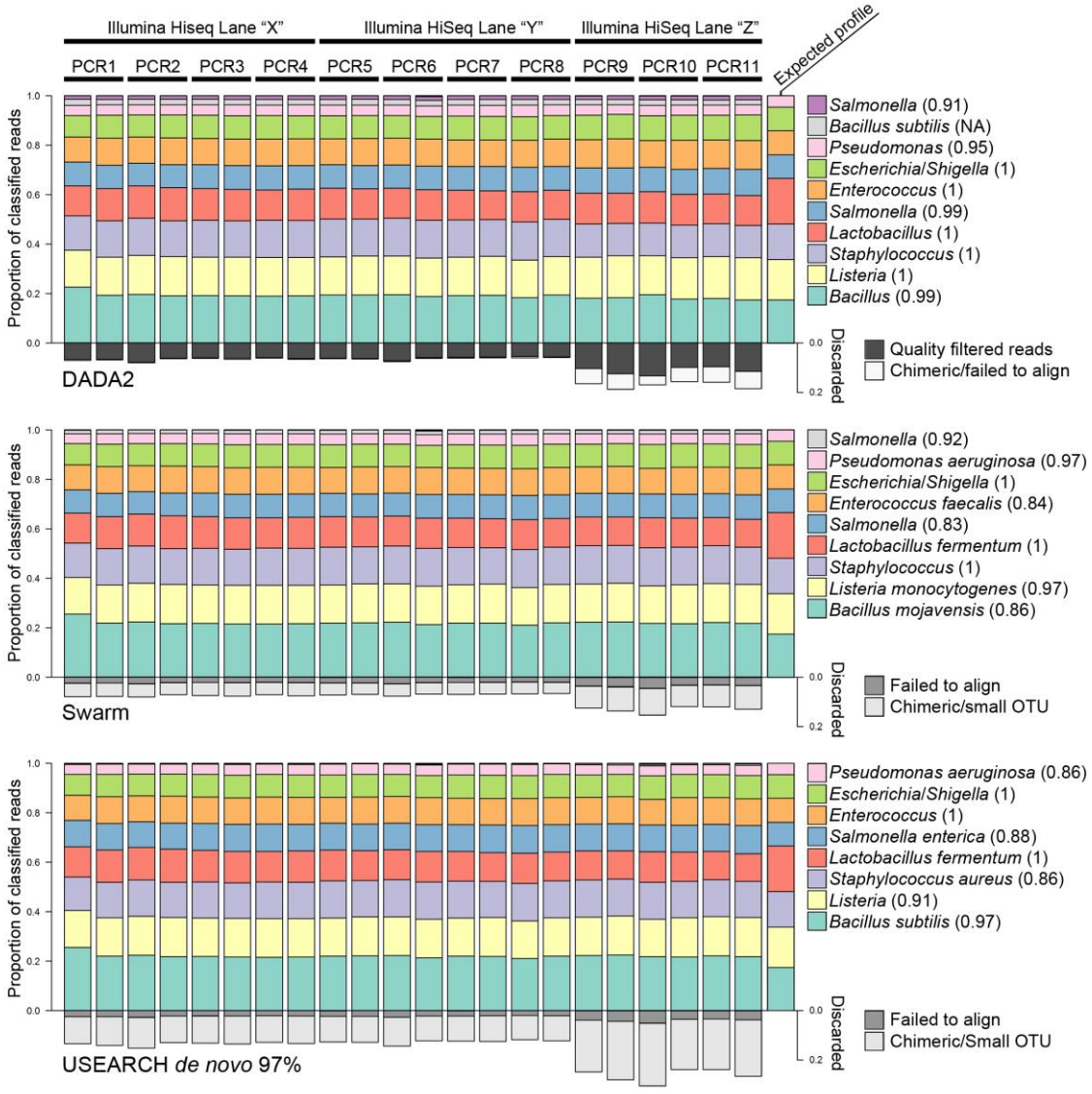


Figure 3.2 Profiles obtained from the Zymo Microbial DNA Standard using six different clustering algorithms. PCR and sequencing run ID is given at the top. OTUs/sequence clusters are coloured according to their taxonomic assignment (see key), with the confidence in the assignment given in brackets. Discarded reads are shown as a proportion of total reads present in the OTU table. The expected profile (which takes account of 16S copy number) is shown at the right side of each chart. Continued on next page.

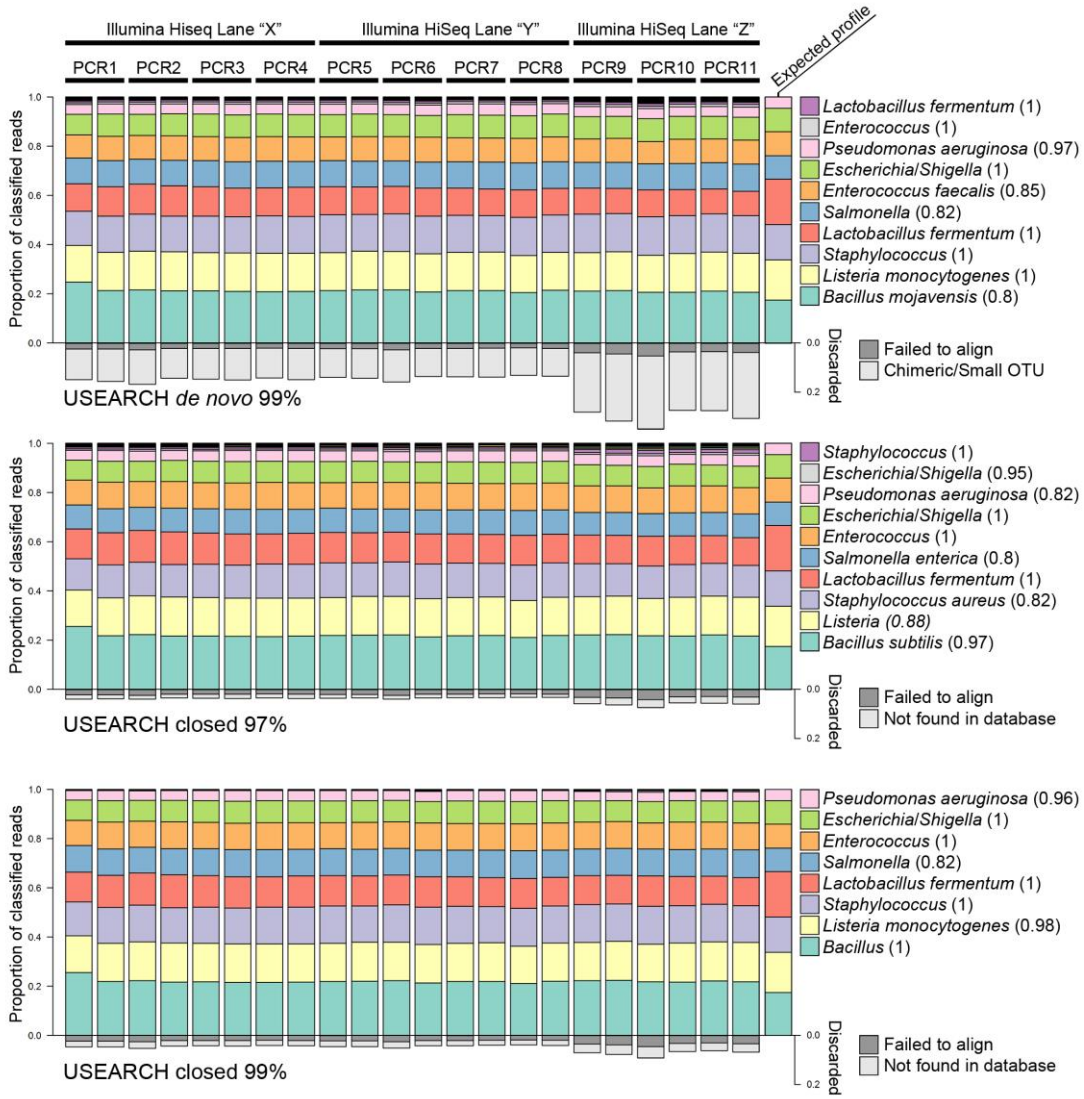


Figure 3.2 (continued)

not generate this cluster. It is difficult to determine whether these clusters reflect a mutation that has occurred in culture, or whether this represents a common sequencing or PCR error. When comparing replicates within a clustering method, the similarity as measured by Bray Curtis score was high for all methods, ranging from 93.4-99.5%, with all methods performing similarly well. Furthermore, similarity scores between the obtained and expected profiles (as determined by shotgun sequencing) were high, ranging from 88.5-93.3%, with the highest similarity scores achieved by DADA2 (90.5-93.3%). It should be noted that this similarity does not take into account the additional smaller sequence clusters generated by Swarm and DADA2, which would increase the similarity slightly for these methods. The differences between methods were significant in a mixed effects linear regression model which controlled for differences due to PCR and sequencing ($P < 0.0001$).

However, the size of the estimated difference due to clustering method was small, ranging from 0.4-1.6% (95% confidence interval 0.2-1.8%) when compared to DADA2. It should be noted that, since accuracy determined in this way is also influenced by PCR and sequencing error and is therefore sample dependent, DADA2 may not always produce the most accurate result. The addition of sequencing run/lane as a fixed effect did not improve the model fit. However, the effect of run/lane on beta diversity as measured by Bray Curtis was significant by PERMANOVA (DADA2: R2 = 0.65, P = 0.001; Swarm: R2 = 0.26, P = 0.001; USEARCH *de novo* 97%: R2 = 0.29, P = 0.001; USEARCH *de novo* 99%: R2 = 0.45, P = 0.001; USEARCH closed 97%: R2 = 0.47, P = 0.001; USEARCH *de novo* 97%: R2 = 0.35, P = 0.001). Pairwise comparisons found that these differences occurred between the two HiSeq runs as well as between the two lanes on the same run (Figure 3.3). The differences were small, with the mean within-lane Bray Curtis similarity being only slightly higher than the mean between-lane Bray Curtis similarity (difference of mean within-lane and between-lane diversity: DADA2: 1.3%; Swarm: 0.3%; USEARCH *de novo* 97%: 0.4%; USEARCH *de novo* 99%: 0.8%; USEARCH closed 97%: 0.7%; USEARCH *de novo* 97%: 0.5%).

A previous study has also reported bias in sequencing data causing variation between runs which was correlated to genomic G+C content (He et al 2010). The authors hypothesised that the differences between runs were caused by differences in run quality with the lower quality run systematically underrepresenting G+C rich taxa. Interestingly, we have also found that run quality can result in bias between runs (see Appendix D), but were unable to correlate this to G+C content. We also found no correlation between genome G+C content and bias in the Zymo Microbial DNA standard replicates (data not shown). Bias between lanes on the same flowcell has also been reported by others (Aird et al 2011).

3.3.3 Single species samples: Species identification and differentiation

For the single-species vaginal bacterial samples, USEARCH failed to distinguish between *L. crispatus* and *L. amylovorus* in all cases (Figure 3.4). Additionally, with *de novo* clustering at 97% similarity threshold, the *L. jensenii* control was split into two large OTUs, one of which also contained the majority of the *L. crispatus* and *L. amylovorus* sequences (Figure 3.4). *L. crispatus* and *L. amylovorus* have over 99.5% sequence similarity in the V3-V4 region and the inability of USEARCH to separate them at 97% or 99% similarity threshold is therefore to be expected. The

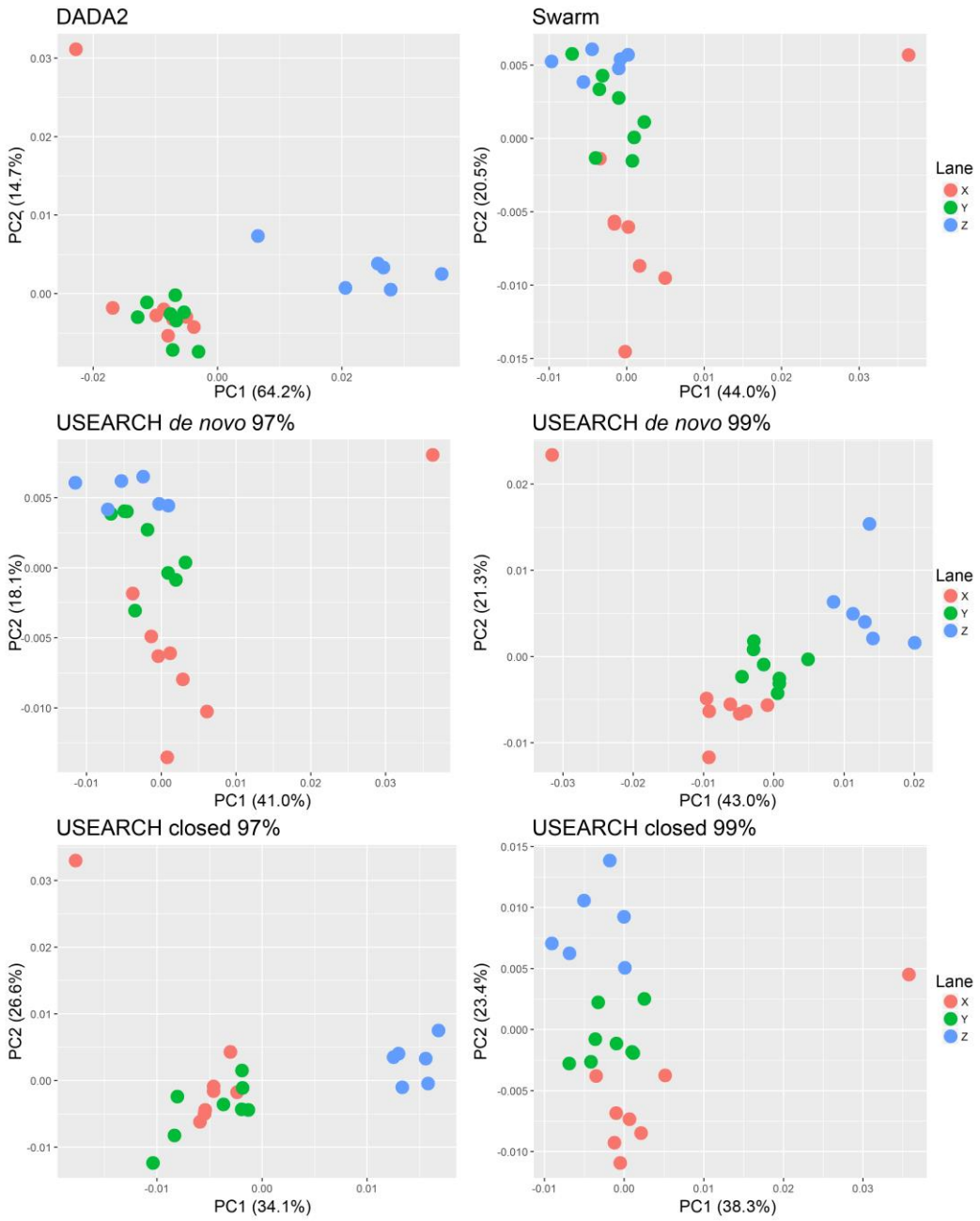


Figure 3.3 Principal coordinates analysis plot of Bray Curtis similarity scores for the Zymo Microbial DNA Standard replicates coloured by HiSeq lane. Lanes X and Y were on the same HiSeq run, while lane Z is on a separate run. Figures in brackets indicate the percentage variation explained by PC1 and PC2.

similarity between *L. crispatus* and *L. jensenii* is lower, around 97% (see Appendix E), with the exact value depending on the sequence variant. The separation of the *L. jensenii* control into two large OTUs has probably occurred because many sequences in that sample were just within the 97% similarity cut-off when compared to the chosen centroid sequence whereas other sequences were just outside this cut-off. It is likely that the selected centroid originated from *L. crispatus* (or the closely related *L. amylovorus*), as this would fit with the *L. jensenii* control being split

into two OTUs, rather than the *L. crispatus* control. USEARCH is input order dependent in that it chooses the first sequence it comes across as centroid (He et al 2015, Mahé et al 2014), if the centroid sequence had come from the *L. jensenii* sample, the controls would likely have looked very different. This "OTU instability" is a major disadvantage, because it makes it potentially difficult to compare between different analyses, even if the exact same bioinformatics pipeline was used (He et al 2015). Similarly, when clustering *de novo* at a 99% similarity threshold, USEARCH produced two large OTUs in both the *L. amylovorus* and *G. vaginalis* controls, most likely for similar reasons (note that in the case of *G. vaginalis*, there were other strains present in the clinical samples which could have affected the clustering of the *Gardnerella* controls, see later). In contrast, OTU picking against a reference database did not result in significant OTU splitting, even though this is theoretically possible.

In contrast to clustering with USEARCH, all species were distinguished by DADA2 and Swarm. However, there was some splitting of samples into several sequence clusters in the case of DADA2. The *Prevotella bivia* sample consisted of three such clusters, with the representative sequences differing from one another by no more than 2 base substitutions. The single sequenced genome of *Prevotella bivia* currently held in the NCBI database (accession NZ_AJVZ00000000.1) is reported to have 4 copies of the 16S rRNA gene. Since the sequence clusters are roughly in the ratio of 2:1:1, the DADA2 profile could reflect true biological variation of 16S rRNA gene copies within the genome. Swarm would be expected to merge these clusters due to their high degree of similarity (d=1). DADA2 also produced a large number of smaller sequence clusters in the *L. iners* and *L. jensenii* controls. These were not present in Swarm clustered profiles, even before filtering of small OTUs. The majority of these were identified as *Lactobacillus* (with only some very small clusters totalling ≤0.08% of the sample identified as other genera). The majority of these were the same length as the main sequence cluster in each control and differed to this by only a single base substitution. The locations of these substitutions occurred throughout the length of the read. Although this could represent biological variation between individual cultured bacteria, it is more likely that this is due to insufficient correction of sequencing (or PCR) error, since many of the sequencing clusters present in one replicate and representing up to 2.3% of that sample are completely absent from the other replicate. The presence of these small clusters will erroneously inflate measures of alpha diversity.

The representative sequences for the main sequence clusters identified by DADA2 and Swarm are identical, with the exception of the *L. iners* controls for which the sequence differs by a single nucleotide. The whole genomes of the bacterial strains used in this study have not been sequenced and it is therefore not possible to determine with certainty which is the "correct" sequence. However, DADA2 identified two sequence clusters matching *L. iners* in sample S14 (see later), one of which is the same as that in the single-species control and the other one matches that identified by Swarm. Since Swarm is expected to merge reads with only one base difference, it is likely that the reference sequence for Swarm derives from the numerically more abundant reads in S14, explaining the single base difference to the reference sequence identified by DADA2. The remaining representative sequences are identical to 16S rRNA sequences in the NCBI database for their respective species. In the case of the four USEARCH datasets, all representative sequences are identical to one another, except where the *Lactobacillus* reads have been subsumed into the same OTU, in which case the representative sequence is the same as for the *L. crispatus* reference obtained by DADA2 and Swarm. This makes sense considering that the majority of sequences within this OTU are from the *L. crispatus* controls (since there is only one *L. amylovorus* control). With the exception of these lactobacilli and *Gardnerella*, the representative sequences obtained by USEARCH are the same as those obtained by Swarm. The representative sequences for the *Gardnerella* controls differ from each other by one insertion/deletion and three base substitutions, but both are identical to sequences held in the NCBI database for this species. However, neither DADA2 nor Swarm identified multiple sequence clusters in the *Gardnerella* controls, as would be expected if both sequences really were present in these controls. The reason for this discrepancy appears to be the high abundance of *Gardnerella* in the S14 controls (see later). This *Gardnerella* strain and that found in the positive control have been placed in separate sequence clusters by both DADA2 and Swarm. In contrast, USEARCH has merged these reads into one OTU, and the more abundant reads from S14 have therefore been assigned as the representative sequence for this OTU. In other words, the representative sequence determined by USEARCH is probably not actually present in the single-species *Gardnerella* controls.

In terms of taxonomic assignments, DADA2 generated the most species-level assignments as oppose to genus-level (Figure 3.4). DADA2 identifies species by exact unique matches to a database, which is based on the assumption that the

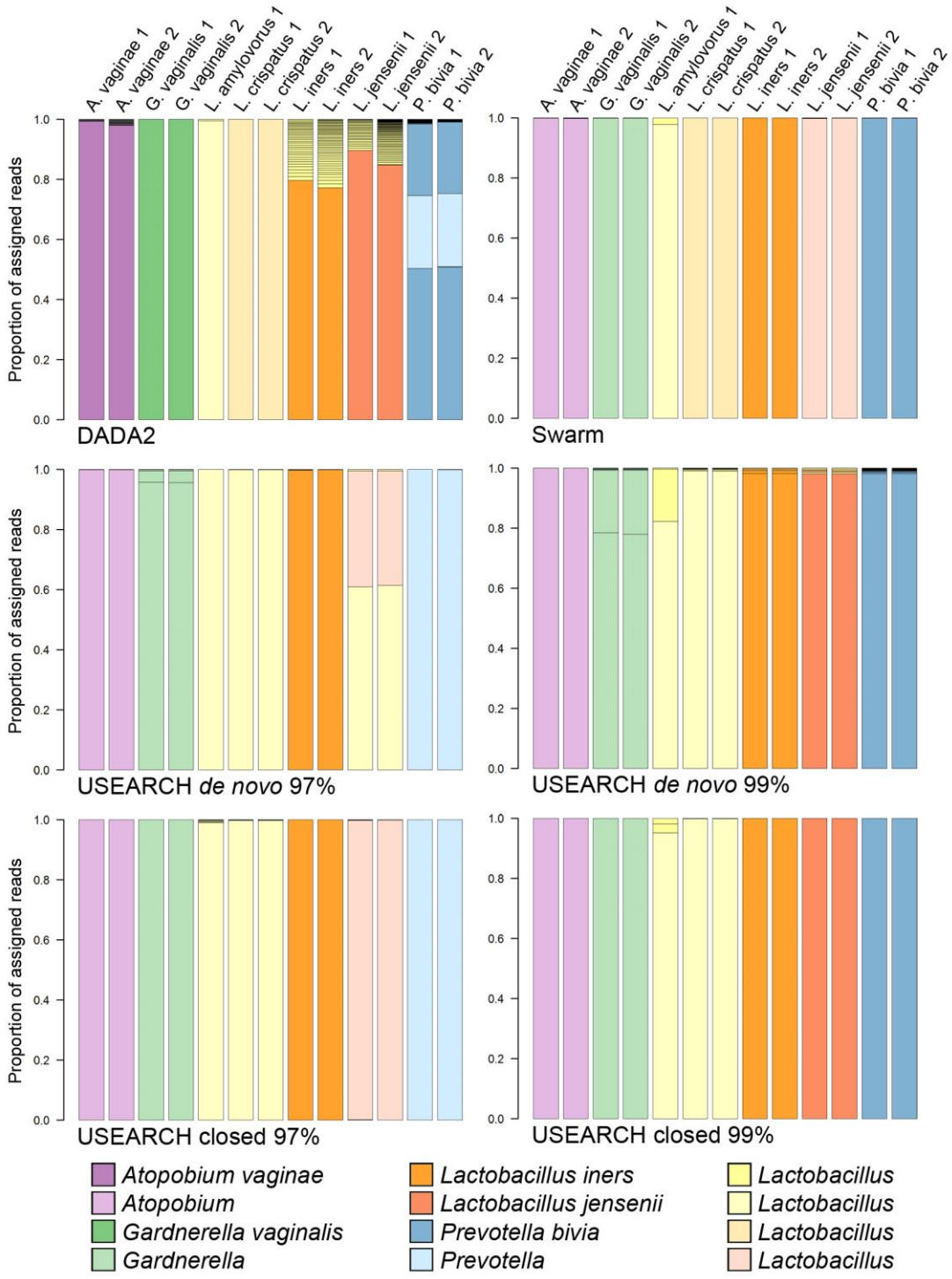


Figure 3.4 Barchart showing profiles of monoculture samples obtained using the six different clustering methods. OTUs/sequence clusters are coloured according to their taxonomical assignment (see key), with dark shades representing species level assignments and lighter shades representing genus level assignments. Several shades were used for *Lactobacillus* OTUs, to allow distinguishing of major *Lactobacillus* OTUs/groups between samples. Both DADA2 and Swarm were able to distinguish between all seven species.

algorithm is able to compute the real DNA sequence by utilising error information. In this way, species that share more than 99% of their DNA sequence can still be accurately differentiated (Callahan et al 2016). As such it is not surprising that this method performed best. Taxonomy was assigned to the other datasets using the Silva database clustered at 97% or 99%. Unsurprisingly, the latter produced more species identifications, but failed to classify *L. jensenii* to species level in the Swarm dataset, even though the representative sequences were identical, which again highlights the stochastic nature of this process (see Figure 3.4). When the 99% clustered database was used to assign taxonomy for the USEARCH datasets clustered with a 97% similarity threshold, there were a similar level of species-level assignments to the USEARCH clustered datasets using a 99% cut-off. However,this will be inaccurate when there are closely related species present in the dataset because the minor species in the OTU will receive the taxonomical assignment determined based on the majority species and is therefore not advisable.

When comparing replicates of the VMB mock community (pooled pre-PCR) within a clustering method, the Bray Curtis similarity was high for all methods, ranging from between 95.4-99.9%. However, although the mock community had been put together to result in approximately equal genome numbers, the proportions of each species were skewed. A degree of dissimilarity is to be expected because there are a variable number of 16S rRNA gene copies per genome in different species (Kembel et al 2012). Although 16S rRNA copy number information is not available for the strains used here, there is information on copy number for these species available in the NCBI database (one copy in *L. iners* and *A. vaginae*, two in *G. vaginalis and four in L. jensenii*, *L. crispatus* and *P. bivia*). However, this alone would not cause differences of the observed magnitude (Figure 3.5). Since DNA was pooled after extraction – assuming genome size estimation was relatively accurate – this bias must have been introduced either during the PCR, by the sequencing or by the bioinformatics pipeline. Bias of this type during the sequencing process is reportedly low (Brooks et al 2015) and is unlikely to have caused the differences seen here since the observed pattern does not reflect the total read count for each of the single-species samples, as might be expected if such bias existed. Bias resulting from the bioinformatics pipeline may occur due to differential read quality from sequences originating from different bacterial species leading to preferential discarding of reads from a particular species (see Appendix D). However, the proportion of discarded reads in these samples was relatively low with

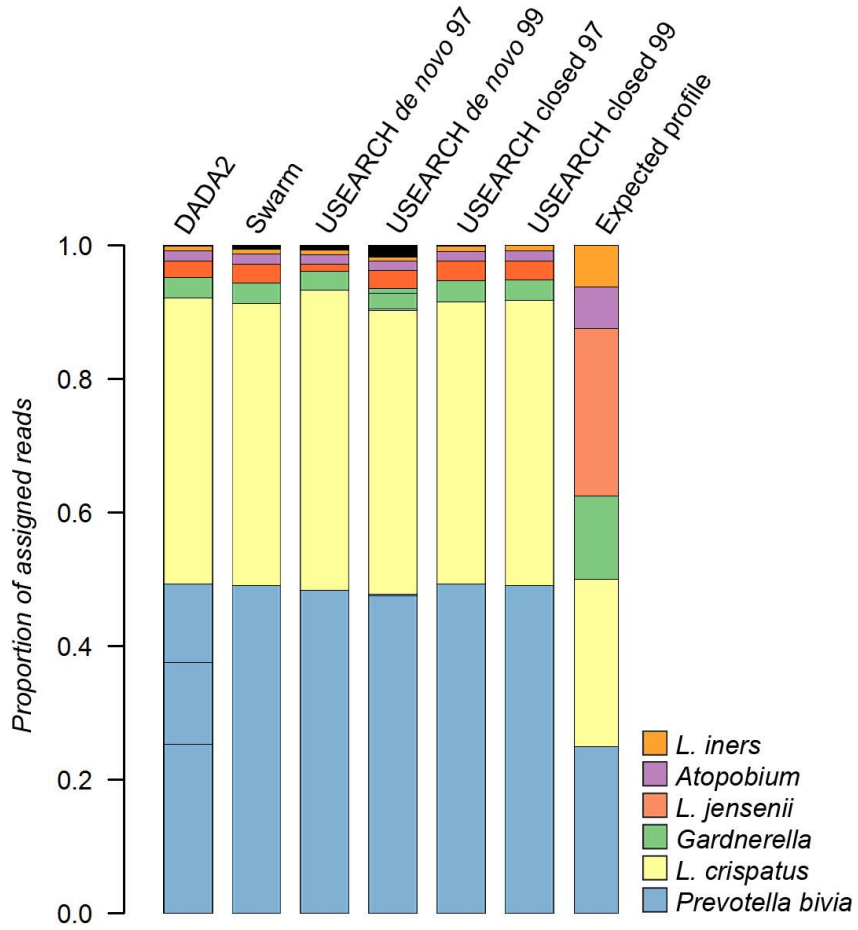


Figure 3.5 Barchart showing microbiome profiles of a single replicate of the VMB mock community (pooled prior to PCR) produced by the different clustering algorithms. The expected profile is based on an estimate of 16S rRNA copy number.

all methods (median 11.5% across all methods), and could therefore not result in differences of this magnitude. Therefore, the most likely source of bias is the PCR reaction, which has resulted in overrepresentation of *L. crispatus* and *P. bivia* and underrepresentation of *L. iners*, *L. jensenii, A. vaginae* and *G. vaginalis*. This was supported by the results of the VMB mock community in which pooling the same six vaginal species was done after PCR, producing profiles that were much closer to those expected, although some differences remained, possibly due to pooling inaccuracies (see Figure 3.6). Previous studies have found that the PCR step (including primer choice) is one of the most significant sources of bias in 16S rRNA amplicon studies (Brooks et al 2015, Hong et al 2009, Schirmer et al 2015, Tremblay et al 2015). It has been suggested that this may be the result of primer mismatches or degeneracies (Tremblay et al 2015), but in this case the sequence in the primer regions was predicted to be the same for all six species (as determined from NCBI database matches to representative sequence). A further hypothetical explanation for PCR bias is that templates with low G+C are preferentially amplified

because they dissociate more efficiency into single-stranded DNA (Suzuki and Giovannoni 1996). This may explain the low proportions of *G. vaginalis* and *A. vaginae* seen as these have a high G+C content (56.8 and 58.5%, respectively) in the amplified region. However, it does not explain the difference in the case of *L. iners* and *L. jensenii* which have a very similar G+C content (50.3 and 51.5%, respectively) to *L. crispatus* and *P. bivia* (51.0 and 50.2%, respectively). Amplicon length does not explain the observed patterns either since this was the same for all *Lactobacillus* species (429 bases), followed by *Prevotella* (424 bases), *Gardnerella* (412 bases) and *Atopobium* (410 bases).

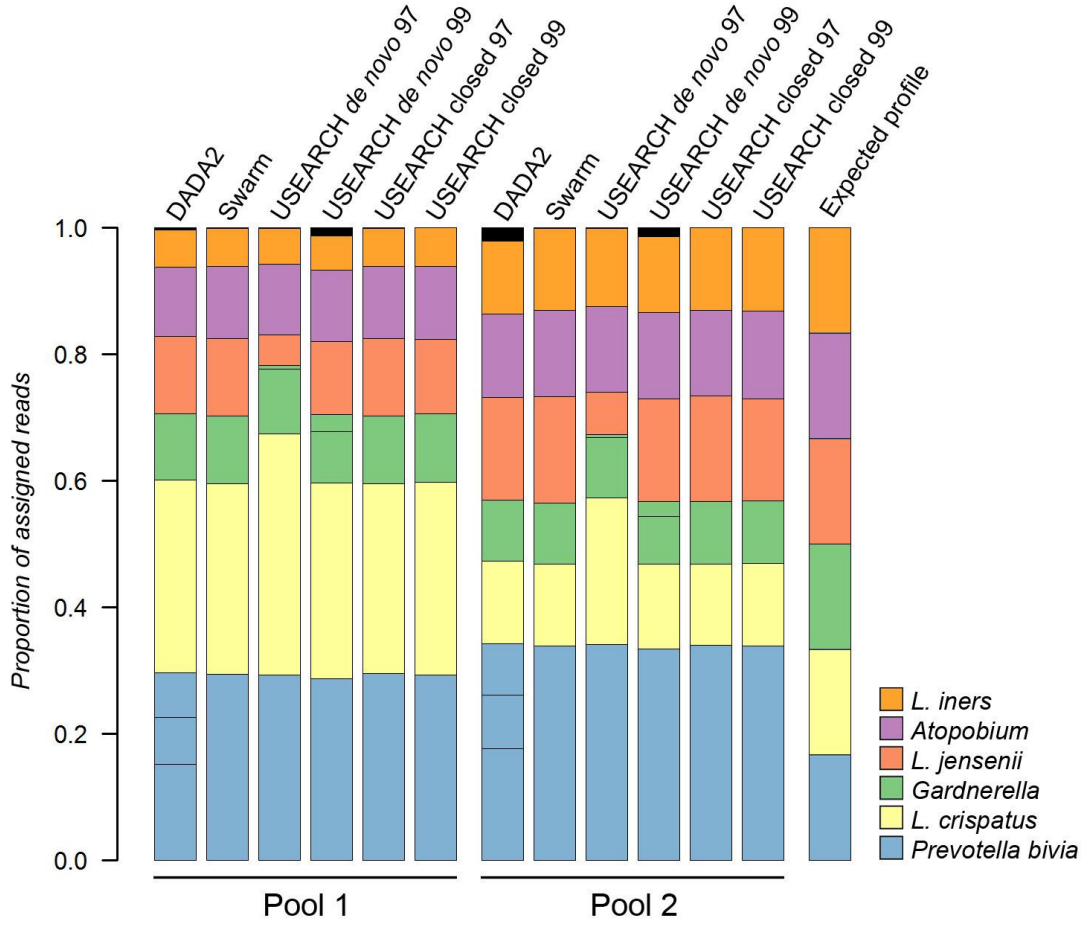


Figure 3.6 Barchart showing microbiome profiles for the VMB mock community pooled post-PCR. Two experimental replicates are shown, each from a different pool and clustered by each of the different clustering algorithms. The three replicates produced from each of the four pools produced similar profiles regardless of clustering algorithm, but there was a significant effect of pool on the profiles obtained (PERMANOVA; P = 0.001 for all clustering methods). The most likely cause for this difference is inaccuracies during sample pooling, which is explained by the low amplicon yields obtained from these controls. A single replicate is shown for each of two of the pools that consistently showed the highest degree of Bray Curtis dissimilarity.

3.3.4 Vaginal samples: Consistency across PCRs and sequencing runs

The main taxa observed in the two vaginal samples were consistent across different methods. S14 consisted mainly of *L. iners* and *G. vaginalis* reads (Figure 3.7) and the reference sequence for the largest cluster for each bacterium (making up 67-83% and 15-23%, respectively) was identical across all methods. With the exception of closed clustering with USEARCH, sample profiles also contained smaller OTUs identified as *L. iners* and *G. vaginalis*. Furthermore, all methods identified sequences from *Lactobacillus* (*vaginalis* and *coleohominis* types) and *Streptococcus* (*pneumoniae* type), which made up no more than 0.74% of any replicate. However, the representative sequences for these smaller OTUs were not necessarily exactly the same. In addition, all methods except DADA2 consistently identified *Ureaplasma* and *Dialister*/*Veillonellaceae*, making up no more than 0.07 and 0.02%, respectively. These taxa were also present in some replicates classified by DADA2 at similar levels, but were completely absent from others.

S11 was a high diversity sample in which all methods detected sequence clusters from *Peptostreptococcus* (*anaerobius*; 19-28%), *Sneathia* (*amnii*; 20-26%), *Prevotella* (*timonensis*; 6-16%), *Mycoplasma* (*hominis*; 5-6%), *Prevotella* (*bivia*; 5-22%), *Streptococcus* (1.6-2.5%), *Aerococcus* (*christensenii*; 0.7-1.2%), *Bacteroides* (*fragilis*; 0.3-0.6%), *Dialister*/*Veillonellaceae* (0.2-0.7%) and *Peptoniphilus* (0.2-0.3%), with the same reference sequence across all methods (Figure 3.7). Note that the species names were mostly assigned only by DADA2. The high degree of variation in relative abundance of *Prevotella* (*timonensis*) and *Prevotella* (*bivia*) is due to their lower relative abundance in the DADA2 classified reads. However, DADA2 detected two and one additional sequence clusters for these species respectively, which together make up a similar relative abundance to other methods. In addition to these taxa, all methods detected a sequence cluster for *Gardnerella* (*vaginalis*; 0.8-1.5%), but with slightly different reference sequences (similarity >99%). As determined by oligotyping, *Gardnerella* is reported to be a highly diverse genus and this may have biological implications (Eren et al 2011). In this case, the reference determined by DADA2 is probably the correct one. This is because DADA2 is the only method that consistently identified a different *Gardnerella* cluster (across all replicates) in the S11 control when compared to both the S14 control and the *Gardnerella* single-species control. The reference sequences for USEARCH likely derive from S14 (see above) and that from Swarm derives from the *Gardnerella* single-species control, which (according to DADA2) differs by only one

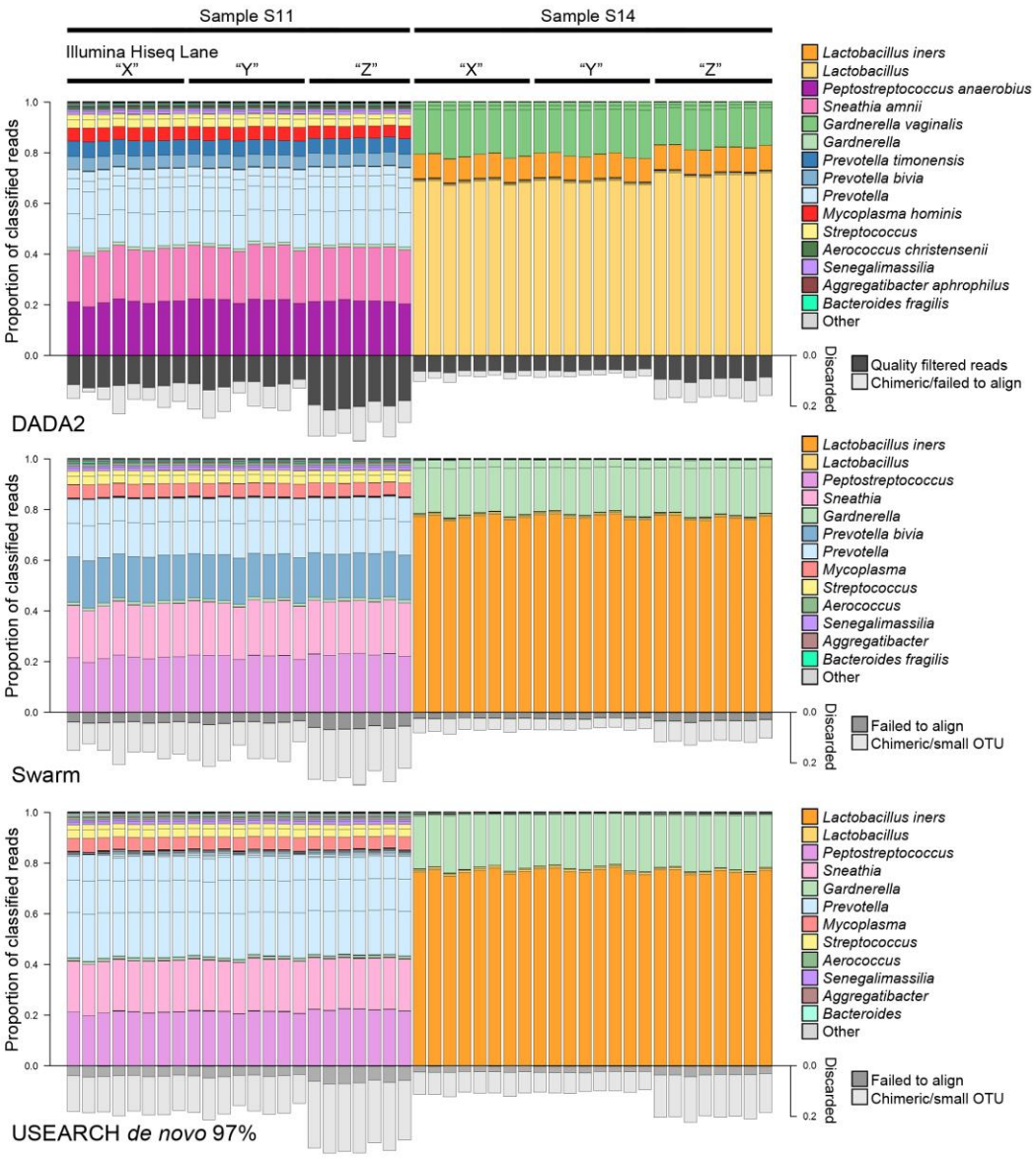


Figure 3.7 Profiles obtained from vaginal samples S11 and S14 using the six different clustering algorithms. Sequencing run ID is given at the top. OTUs/sequence clusters are coloured according to their taxonomic assignment (see key), with dark shades representing species level assignments and lighter shades representing genus level assignments. Discarded reads are shown as a proportion of total reads present in the OTU table. Continued on next page.

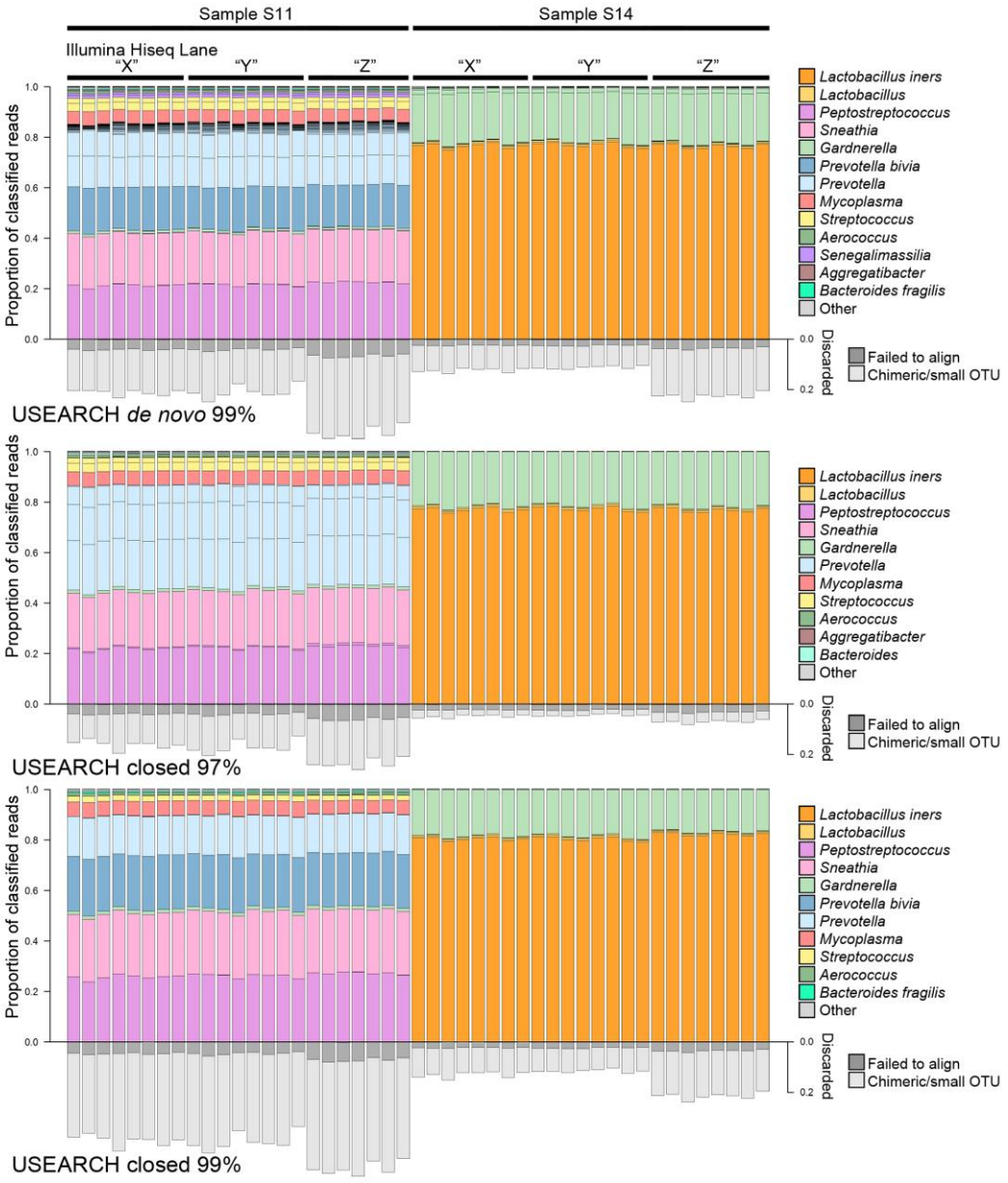


Figure 3.7 (continued)

base substitution to that in S11. An inability to separate sequences differing by only one base is one of the limitations of Swarm. Taxa detected only by the *de novo* methods were *Senegalimassilia* (1.1-1.3%), *Aggregatibacter* (*aphrophilus*; 0.34-0.53%), *Howardella* (0.09-0.16%) and *Paracoccus* (0.03-0.07%), as these taxa are not contained in the vaginal database. However, they have been isolated from other human body sites (Cundell 2016, Lagier et al 2013, Rylev et al 2011, Yang et al 2015) and could therefore represent genuine members of the microbiota. *Paracoccus* has also been reported as a reagent contaminant (Salter et al 2014), but was not present in the negative extraction controls in this study. The reference sequences for these bacteria were identical for all methods. A further four sequence

clusters were detected by all methods except closed USEARCH at 99%: *Streptococcus* (2.8-3.5%), two *Prevotella* clusters (5-11% and 0.16-0.37%), and *Fusobacterium* (0.002-0.09%) presumably because the sequence similarity threshold of 99% was too stringent to capture all (or most) of the reads when clustered against the vaginal database. The reference sequences were identical for all of these taxa, except *Fusobacterium* for which Swarm, DADA2 and USEARCH (*de novo* at 99%) identified the same sequence and the remaining USEARCH methods (*de novo* and closed at 97%) identified another. Although these shared only 97% sequence similarity with one another, both sequences have identical matches in the NCBI database. In addition to these taxa, Swarm consistently identified *Anaerococcus* (0.004-0.017%), *Olsenella* (0.003-0.015%), *L. iners* (0.001-0.012%), *Bulleidia* (0.001-0.008%) and *Finegoldia* (0.001-0.008%) in the S11 control. These same taxa were also identified by DADA2 (with the same reference sequences except for *Bulleidia* which differed by one base substitution), but inconsistently so. Some of these were also detected consistently by USEARCH: *Anaerococcus*, *L. iners*, *Bulleidia* (only with similarity threshold of 97%) and *Finegoldia* (only with closed). Since these were low abundance taxa, they may have been absent from some of the USEARCH clustered datasets due to the size cut-off used for filtering small OTUs. All methods, but in particular Swarm and USEARCH *de novo* contained additional small sequence clusters that were identified as the same taxa as larger clusters already mentioned. These may represent sequence variants due to error or less likely, biological variation. In the case of Swarm and USEARCH *de novo* (97% similarity), the relative abundance of these clusters was no greater than <0.2%, but was higher for Usearch *de novo* (99% similarity), which makes sense since the variation in a given sequence cluster is more limited and increased numbers of reads that would otherwise have been assigned to a larger OTU are assigned to a different OTU as they lie beyond the 99% cut-off. In order to reduce the number of OTUs that represent sequencing error, a size cut-off was employed to filter small OTUs from the dataset. This is based on the assumption that small OTUs are more likely to represent read error than real biological variation and is thought to greatly improve estimates of diversity (Bokulich et al 2013). Even though this strategy risks filtering out rare taxa which can have significant ecological importance (Lynch and Neufeld 2015), it is interesting to note that the consistency of detection of small OTUs was superior to that of DADA2 which employs a different strategy in which filtering of small sequence clusters is not applied.

In general the consistency across methods for each of the two vaginal samples was high with similar results obtained for all methods. Bray Curtis similarity ranged from 93.0-99.7% for S11 and 94.3-99.9% for S14 (the slightly higher similarity between replicates of S14 is expected since this sample was lower diversity). However, in some cases there is a small but discernible difference between replicates from different HiSeq runs. This is most noticeable with the results for S14 clustered by DADA2 and USEARCH closed at 99% similarity, where the replicates from run Z have a slightly higher proportion of the *L. iners* sequence cluster and a lower proportion of the *G. vaginalis* sequence cluster, compared with the lanes from the other sequencing run (Figure 3.7). This difference was significant by PERMANOVA (DADA2: R2 = 0.76, P = 0.001; USEARCH: R2 = 0.60, P = 0.001). Pairwise comparisons showed that this difference was due to variation between the two HiSeq runs rather than between the two lanes from the same run. The difference between mean within-lane and between-lane diversity was 1.5% for DADA2 and 0.7% for USEARCH. With the other methods, no difference could be discerned for this sample (Figure 3.8) and the mean within-lane beta diversity was no greater than the between-lane diversity. In the case of DADA2, the relative abundance shift was more marked in an earlier analysis that had used more stringent quality cut-offs. These differences may be related to differential loss of sequencing reads that did not pass the quality cut-off. Run Z had an earlier quality drop-off in read 2 compared to the two lanes from the other run, which resulted in comparatively higher read loss due to quality filtering (Figure 3.7). Since read loss can be biased towards certain species (see Appendix D), this could have caused the differences between runs. The reason for the difference seen with USEARCH is unknown, since the quality filtering used was the same as for the Swarm and other USEARCH pipelines. Additionally, the percentage of raw reads retained in the final OTU table was comparable to USEARCH *de novo* clustering at 99% cut-off. Sample S11 is of higher diversity, making it more difficult to visualise differences between runs. However, there are differences between runs observed on PCoA plots for all methods (Figure 3.9), which are statistically significant by PERMANOVA (DADA2: R2 = 0.33, P = 0.002; Swarm: R2 = 0.32, P = 0.008; USEARCH *de novo* 97%: R2 = 0.37, P = 0.003; USEARCH *de novo* 99%: R2 = 0.43, P = 0.001; USEARCH closed 97%: R2 = 0.58, P = 0.001; USEARCH *de novo* 97%: R2 = 0.39, P = 0.002).

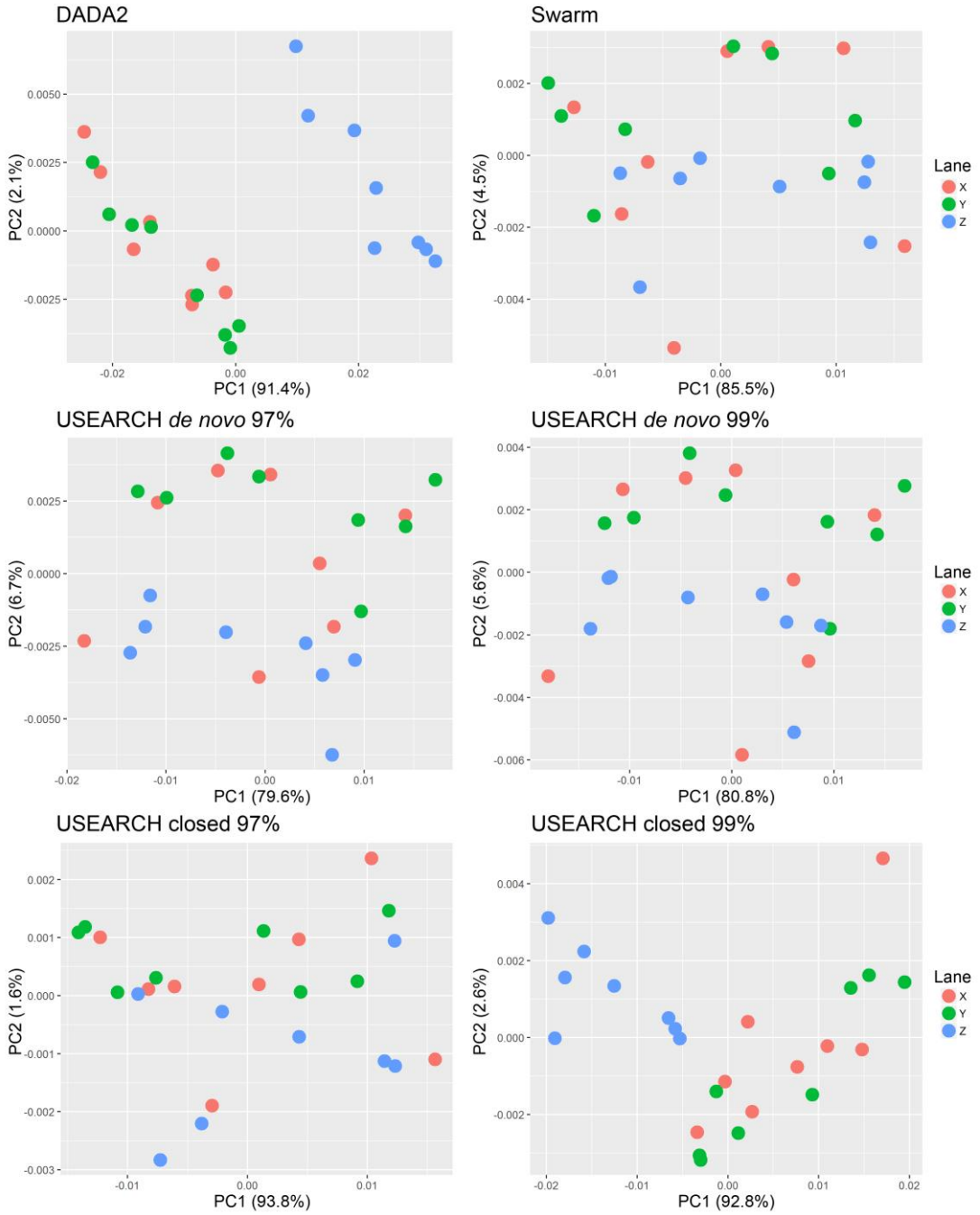


Figure 3.8 Principal coordinates analysis plot of Bray Curtis similarity scores for replicates of sample S14 coloured by HiSeq lane. Lanes X and Y were on the same HiSeq run, while lane Z is on a separate run. Figures in brackets indicate the percentage variation explained by PC1 and PC2.

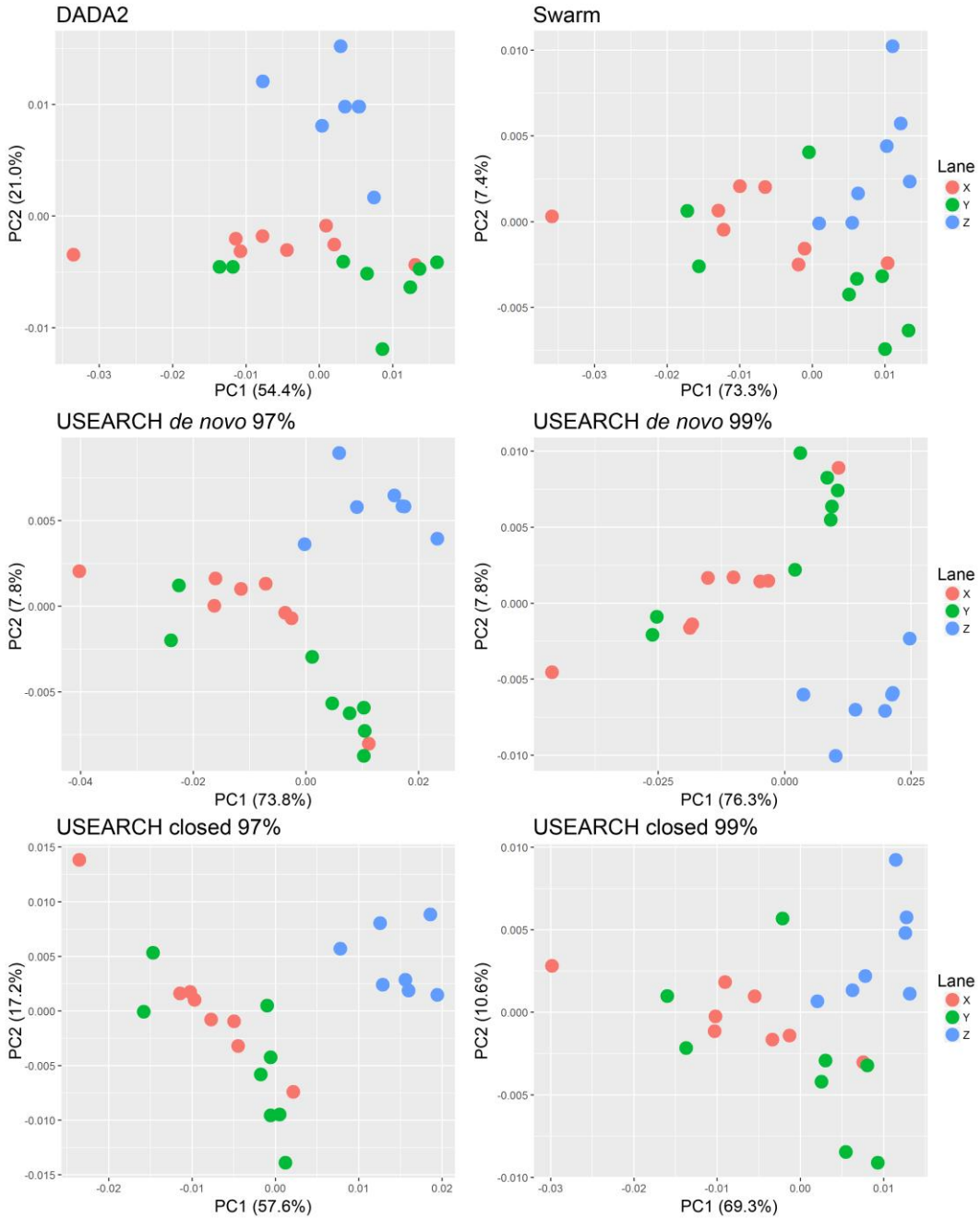


Figure 3.9 Principal coordinates analysis plot of Bray Curtis similarity scores for replicates of sample S11 coloured by HiSeq lane. Lanes X and Y were on the same HiSeq run, while lane Z is on a separate run. Figures in brackets indicate the percentage variation explained by PC1 and PC2.

Pairwise comparisons showed that this difference was due to variation between the two HiSeq runs rather than between the two lanes from the same run. These differences were again relatively small (difference of mean within-lane and between-lane diversity: DADA2: 0.4%; Swarm: 0.3%; USEARCH *de novo* 97%: 0.6%; USEARCH *de novo* 99%: 0.8%; USEARCH closed 97%: 0.8%; USEARCH *de novo* 97%: 0.4%).

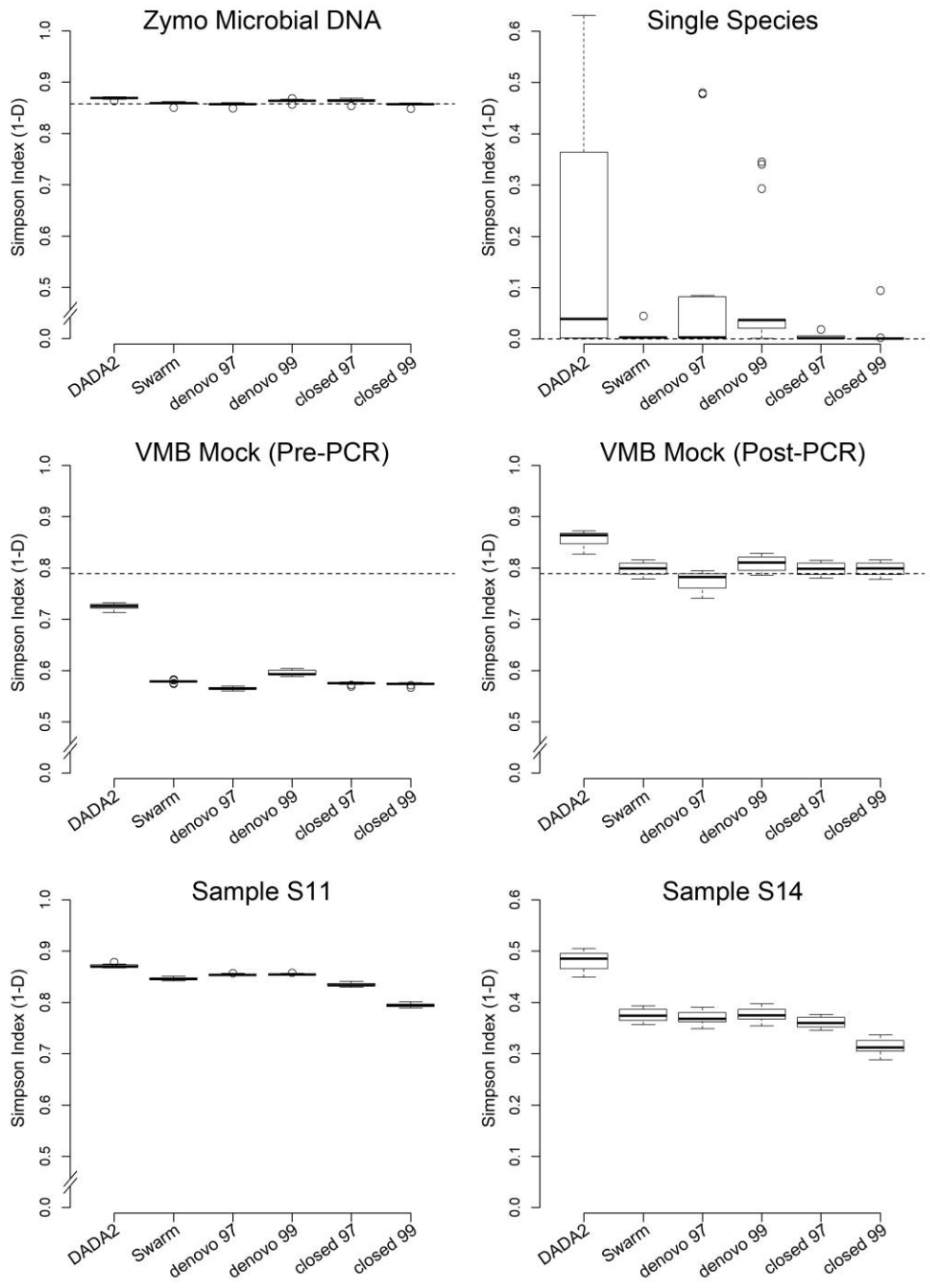


Figure 3.10 Alpha diversity analysis of control samples using different clustering algorithms

3.3.5 Effect on alpha diversity

The Simpson index (1-D) was used as a measure of alpha diversity as this is relatively robust to changes in the numbers of small OTUs and therefore provides a fair comparison that is not affected by small scale errors such as barcode switching and variable cut-offs for discarding of small OTUs (Figure 3.10). DADA2 tended to produce the highest diversity estimates and, where the expected diversity was known, usually overestimated diversity. The exception to this was the VMB mock community that had been pooled prior to PCR, for which all methods

underestimated alpha diversity due to the skewing of the relative abundance of species caused by PCR bias (see section 3.3.3). While the diversity index for the DADA2 profiles is closest to that expected, this can be at least partially explained by the splitting of *P. bivia* into three separate clusters (see section 3.3.3) which would have compensated somewhat for the PCR bias. This splitting also increased the diversity estimates of the single species samples clustered with DADA2, where not only the *P. bivia* samples (0.63 in both cases), but also the *L. iners* (0.36 and 0.40) and *L. jensenii* samples (0.20 and 0.28) had much higher diversity indices. In the case of *P. bivia* this increased diversity is perhaps not biologically correct since there is only a single species present in the sample. However it makes computational sense. On the other hand, the increased diversity estimates of the *L. iners* and *L. jensenii* samples have occurred due to the presence of high numbers of presumably erroneous OTUs.

Using reference-based (closed) OTU-picking with USEARCH, the diversity of the vaginal samples was lower compared with the other clustering methods, particularly using a cut-off of 99%. While it is possible that this is due to fewer erroneous OTUs being identified, it is more probable that some OTUs are missed due to their absence in the database.

3.4 Discussion

As the use of 16S rRNA amplicon studies has become more widespread, there has been increased interest in experimental error resulting from this technique. One area that has received particular interest is the use of negative extraction controls to identify reagent contaminants, allowing subsequent identification and removal of these sequences from microbiome profiles (Salter et al 2014) and this is rapidly becoming the norm. However, more recently the use of positive controls (e.g. mock communities of known composition) is also being advocated in order to be able to better understand bias resulting from library preparation, sequencing and bioinformatics (Brooks et al 2015, D'Amore et al 2016). These samples can also aid in the optimal choice of methods which may differ depending on the community being studied and the aims of the research. While it is accepted that a degree of bias in 16S rRNA amplicon studies cannot be avoided, an understanding of the biases that are likely to be present in the data is key to accurate interpretation of results. As is evident from the results seen here, different types of controls can be used to answer different questions and the choice of positive controls should

therefore be carefully thought through when designing any 16S rRNA amplicon study. For example, in this study, only the closely related single-species controls were able to detect the shortcomings of USEARCH in classifying closely related taxa.

The ability to distinguish between closely related species is important when characterising the VMB since different species within the *Lactobacillus* genus are known to differ in their associations with health outcomes (Verstraelen et al 2009). In this context, the performance of USEARCH was poor, because it failed to separate closely related species. Additionally, USEARCH *de novo* merged reads originating from different species incompletely, resulting in multiple OTUs in some of the single-species controls, which would confuse downstream analyses. This effect was evident with both similarity thresholds tested (i.e. 97% and 99%). He and others (He et al 2015) recently reported that *de novo* OTU clustering suffered from OTU instability, whilst closed OTU picking did not. The reason for this appears to be the input order dependence of *de novo* clustering with USEARCH which results in a different centroid sequence being chosen with each new input ordering. Whilst closed OTU picking is reported to be stable, this is only the case if the same reference database is used with each iteration, since the stability is affected by the order of sequences in the database (Westcott and Schloss 2015). Furthermore, one recent study found that increased OTU stability did not necessarily translate to increased OTU accuracy as defined by the actual distances between sequences meeting the specified cut-off (Westcott and Schloss 2015). However, the methodology of this comparison is perhaps questionable when considering that a 97% similarity threshold could legitimately result in a pairwise sequence similarity as low as 94% between any two sequences within the same OTU cluster.

In contrast to USEARCH, both Swarm and DADA2 were able to differentiate all seven species tested. However, DADA2 generated a large amount of small sequencing clusters in two of the positive controls, which increased alpha diversity measurements and are likely erroneous. Since the reference sequences for these clusters were all highly similar, USEARCH would most likely have put all of these reads into the same cluster and Swarm may also have done so if the differences between reads were small enough. Alternatively, these errors may have been removed by error correction with SPAdes which was not performed in the DADA2 pipeline as this would have interfered with the error rate inference. The methodology

behind Swarm appears to deal effectively with sequencing error in all the controls analysed here. However, it does mean that Swarm is unable to differentiate between sequences that differ by one base, which was evident for example in the case of *Gardnerella* in sample S11 and the single-species *Gardnerella* control, which was only separated into distinct clusters by DADA2. Of the tested algorithms only DADA2 achieved maximum possible resolution. It should be noted that whilst this increased resolution allows the identification of subtle differences in the DNA sequence data, it does not necessarily correlate perfectly with phylogeny or phenotypic characteristics (Berry et al 2017). One could argue that this high degree of resolution, while computationally correct, may lead to biologically misleading results. For example, DADA2 detected three sequence clusters in the *Prevotella* control, which are likely present within the same genome and this led to an inflated estimate of alpha diversity. It would therefore be difficult to directly compare alpha diversity measures between datasets that have been created using different sequence clustering algorithms. However, this improved degree of resolution does allow for the best possible accuracy in taxonomical assignments, which is of great advantage when studying closely related species, such as the lactobacilli in the vaginal niche. A further advantage of this high resolution is that the obtained sequence clusters are likely to be more consistent and comparable between different studies (Callahan et al 2017).

All methods used in this study were able to detect all species expected to be present in both the VMB and Zymo mock communities. However, both Swarm and DADA2 detected additional sequences in the Zymo standard. A recent study comparing three sequence clustering algorithms, including DADA2, on amplicon data from the V4-V5 region also found that DADA2 detected all expected bacterial sequences with 100% sequence similarity, and additionally found three further sequences that were within 97% sequence similarity to the expected sequences (Nearing et al 2018). However, contrary to our results, the expected ratios of bacterial species were somewhat skewed, which may be due to the use of a different region of the 16S rRNA gene. Interestingly, the authors also concluded that of the methods tested, DADA2 was better at detecting very rare organisms, while in our study DADA2 failed to detect rare species in some replicates. However, this difference may be explained that the two studies did not compare DADA2 to the same algorithms.

With the exception of clusters that were insufficiently separated, the representative sequence obtained by all methods was highly accurate. In the case of DADA2 – which relies on identical sequence matches to call species – this accuracy meant that many of the representative sequences could be assigned to the species level. This is in accordance with the findings of Kopylova and others (2016), who reported that both Swarm and USEARCH 6.1 achieved a high degree of taxonomic accuracy for simulated datasets, but also reported that Swarm outperformed USEARCH 6.1 in this regard on mock communities. In our study, only unique matches were used for species assignments, but DADA2 also allows identification of multiple species with the same sequence, which can be very useful in the vaginal niche where some species share the same sequence over the V3-V4 region, but can still be separated into species groups (see Appendix E). The other methods were similarly accurate in determining the expected representative sequence and it should therefore be possible to use the DADA2 species identification algorithm on the representative set to classify reads to species level. However, this approach falls down with USEARCH when similar taxa are present because they may be merged into one OTU. This is likely to be much less of an issue with Swarm, as representative sequences are unlikely to differ by more than a single base from the true sequence.

In addition to allowing testing of various clustering methods, the positive controls used in this study also allowed the identification of additional sources of bias. This includes an effect of the sequencing run/lane on the observed beta-diversity. The exact mechanism for this is unknown, but may relate to differences in error profiles. DADA2 appeared to be particularly affected by this, despite the fact that it did not have the highest amount of discarded reads compared to the other methods. This suggests, that it may relate to the way that DADA2 uses error data to classify reads. Comparatively speaking, the differences in beta diversity seen here were very small and are unlikely to be biologically meaningful. Nevertheless, this potential for bias should be borne in mind in the experimental design and interpretation of 16S rRNA amplicon studies.

A more significant source of bias was identified by the vaginal mock community, which showed that the PCR reaction resulted in bias by an unknown mechanism. This is a common source of bias in 16S rRNA studies and cannot be avoided with currently available technologies. However, it is reassuring that – while taxa were not in the expected ratio – the community profiles did contain all six vaginal species as

expected. Additionally, while this bias affects measures of alpha diversity, beta diversity measures comparing samples processed in the same way appear to be fairly robust to this type of bias (D'Amore et al 2016, Tremblay et al 2015) and clustering of samples into community types is therefore unlikely to be affected.

3.5 Conclusion

In conclusion, both USEARCH *de novo* and closed are unable to accurately differentiate species if they share a high degree of sequence similarity. This is of particular concern in the vaginal niche where accurate differentiation of the lactobacilli is of particular interest. USEARCH *de novo* suffers from the additional problem of splitting sequences originating from the same bacterial strain into several OTUs if the chosen centroid belongs to a closely related but different bacterium. Although this was not observed with USEARCH closed, OTU splitting could theoretically occur in the same way. In addition, any bacteria not present in the database are missed with this approach, preventing researchers from taking advantage of the ability of sequencing to detect novel species. However, USEARCH closed is extremely fast and as such is highly suited to providing a quick initial look at the results of 16S rRNA sequencing studies.

Table 3.2 Summary of key points regarding different sequence clustering methods. Points in bold represent conclusions from this study.

DADA2 • **Accurate species identification** • **May miss very rare sequence clusters** • **May erroneously inflate alpha diversity** • **Can differentiate single base differences** • **Easy to follow pipeline** • **User defined parameters may affect how sensitive pipeline is to differences in quality between runs**	Swarm • **Deals effectively with read error** • **Can differentiate 2 or more base differences**
USEARCH *de novo* • **Merges separate species if sequence similarity is within similarity threshold** • May split a single species if sequence similarity is without similarity threshold • **May split single species into more than one OTU if closely related species present in dataset**	USEARCH closed • **Merges separate species if sequence similarity is within similarity threshold** • May split a single species if sequence similarity is without similarity threshold • **Unable to detect species not in database** • **Fast**

In contrast, both DADA2 and Swarm were highly efficient at differentiating closely related species. However, DADA2 may miss very rare sequence clusters and may

erroneously inflate alpha diversity of some samples. Neither of these problems was seen with Swarm. While Swarm is unable to distinguish single base differences, this property could be what prevents the alpha diversity inflation seen with DADA2. Considering these results, we will use the Swarm clustering pipeline in the analyses described in the following chapters. In addition, we will use the species identification algorithm implemented in DADA2 on the reference sequences obtained by Swarm in order to achieve more accurate taxonomical assignments. Table 3.2 summarises the key points relating to each clustering method tested.

CHAPTER 4: The VMB-HARP Study

4.1 Introduction

The human microbiome is increasingly being studied as a potential factor in the pathogenesis of infectious diseases and cancer. As previously described, the cervicovaginal microbiome has been implicated as a contributing factor in the acquisition of high-risk human papillomavirus (HR-HPV) infection and the progression to cervical cancer (see section 1.3.4). The increasing affordability of molecular techniques has allowed much more detailed investigation of the microbiome in these states, but the results of currently published studies are often inconclusive and sometimes contradictory.

Between 2011 and 2013 the HPV in Africa Research Partnership (HARP), coordinated by the London School of Hygiene and Tropical Medicine, carried out an epidemiological study designed to improve the early detection rates and treatment for cervical cancer in Africa. During the HARP study, 1250 HIV-positive women were recruited at two clinical centres in Burkina Faso and South Africa to determine HR-HPV infection status and the presence of precancerous cervical lesions by histology. Women without high grade lesions were recalled a median of 16 months later for repeat examination. At both time points vaginal swabs were collected, providing a unique opportunity to study the vaginal microbiome (VMB) in this large cohort of women. The results of this sub-study (the VMB-HARP study) which utilised samples from the South Africa site are described in this chapter.

The HARP study was deliberately limited to HIV-positive women due to their increased risk of persistent HR-HPV infection and the associated increased risk of developing cervical cancer (see section 1.4). Furthermore, HIV infection is particularly prevalent in sub-Saharan Africa (especially in South Africa), making an investigation into the health of HIV positive women particularly pertinent within this region. With relevance to the VMB-HARP sub-study, there is some evidence that HIV infection may modify the relationship between the VMB and HR-HPV infection (Dareng et al 2016) which means that taking account of HIV infection status is vital for the correct interpretation of any studies investigating this relationship. By including only HIV-positive women, the VMB-HARP study ensures that this factor is controlled for, whilst still having a good sample size to detect any associations between the VMB and HR-HPV infection in this group of women.

Previous studies investigating the association of the VMB with cervical cancer have either relied partially or wholly on Pap smear results (Audirac-Chalifour et al 2016, Mitra et al 2015, Oh et al 2015) or excluded women with normal cytology results (Piyathilake et al 2016). The VMB-HARP study benefits from the fact that all women in the parent study underwent four-quadrant cervical biopsies to determine cervical intraepithelial neoplasia (CIN) status, unless they were very unlikely to have any cervical lesions (see section 4.2.6 for further details). All samples classified as high grade (i.e. CIN2 and 3, referred to in this thesis as "CIN2+") and a proportion of those classified as low grade (CIN1) also underwent a consensus review to maximise accuracy of the histology results.

The overall aim of the VMB-HARP sub-study was to determine the association between the VMB and both HR-HPV infection and cervical cancer in HIV-positive South African women. This was achieved by characterising the VMB in women enrolled in the HARP study using 16S rRNA sequencing and the laboratory and bioinformatics methods developed in the preceding chapters.

4.2 Methods: HPV in Africa Research Partnership (HARP)

Samples for the study presented in this chapter were obtained by the HARP team, which I was not a part of. The aims and methodology of the HARP study have been published elsewhere (Kelly et al 2017), but are summarised here where relevant to the VMB-HARP study. The overall aim of the HARP study was to improve cervical cancer prevention programs for HIV–positive women in Africa by evaluating multiple screening methods for effectiveness and thereby allowing the development of cost-effective strategies for earlier detection and treatment. The study enrolled HIV-positive women (N=624 at the South Africa site), recruiting two separate cohorts of women either on antiretroviral therapy (ART) for HIV and those not yet receiving ART, in a ratio of approximately 2:1. Each participant was screened for human papillomavirus (HPV) and cervical lesions at baseline. Women without significant lesions, defined as CIN1 or lower, were re-examined a median of 16 months later to detect incident cervical lesions and persistence, incidence and clearance of HR-HPV. The HARP study also recorded various socio-demographic variables and HIV-related factors (e.g. ART status, plasma HIV viral load, CD4+ counts), and screened for concurrent genital and sexually transmitted infections (STIs).

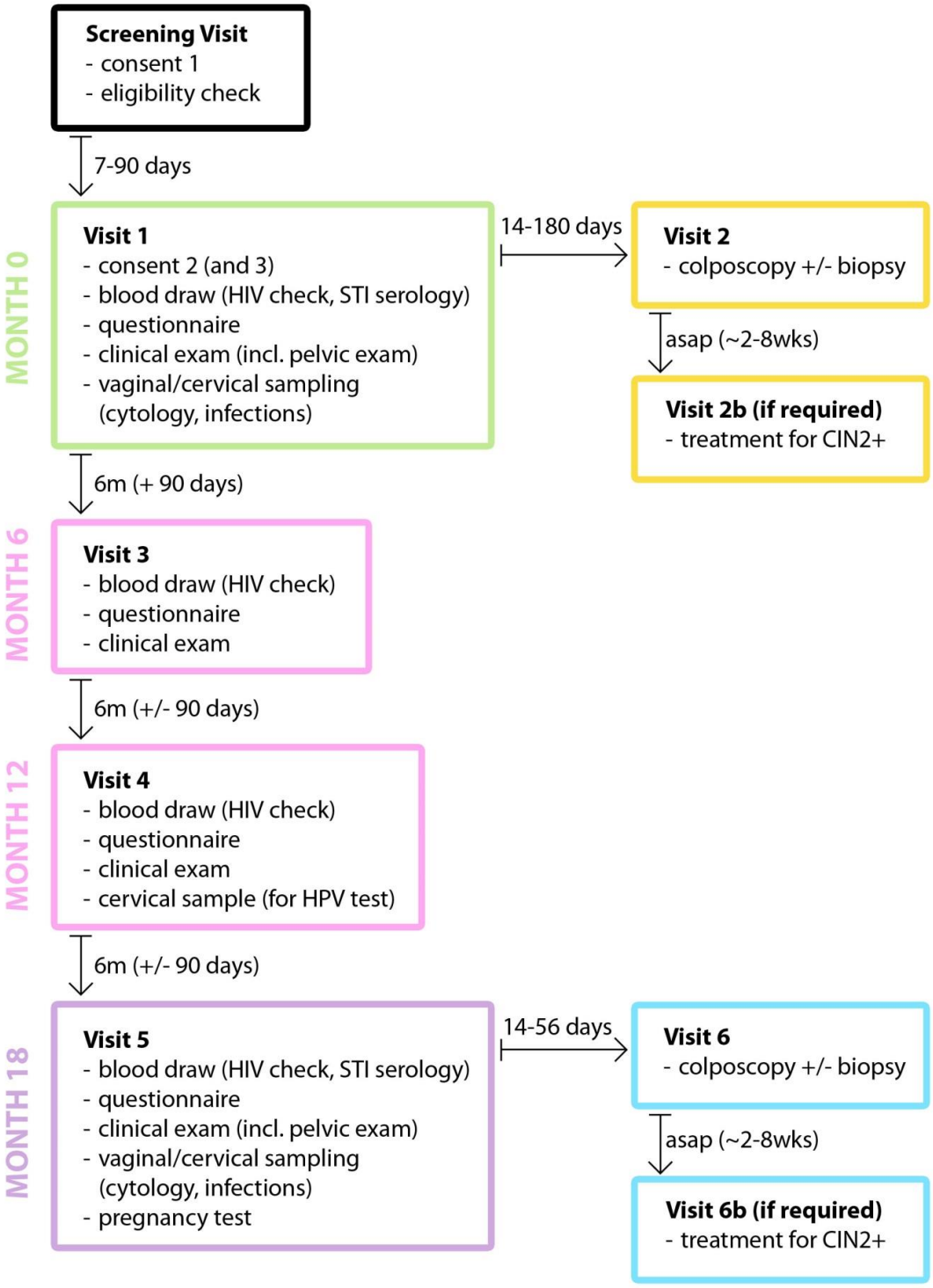


Figure 4.1 Summary diagram of HARP study visits. Samples collected at visit 1 and 5 were used in the VMB-HARP sub-study.

4.2.1 Study population

The South Africa cohort of the HARP study was recruited at the University of Witwatersrand Reproductive Health and HIV Institute (WRHI) research clinic in Johannesburg. Due to cost and logistical reasons, the VMB-HARP study utilised

only samples collected at this site. Women were included in the study if the following criteria were met:

- HIV-1 positive (this was determined by HIV serology or alternatively by the participant providing reliable proof of status)
- 25-50 years of age (inclusive)
- Resident in Johannesburg with no intention of moving in the next 12 months
- No history of cervical cancer or hysterectomy
- Not pregnant at enrolment, not given birth in the last 8 weeks, and not planning to get pregnant in the next 6 months

4.2.2 Participant management

Women that were enrolled in the HARP study attended a number of visits (see Figure 4.1). The planned timings of these visits are given in Figure 4.1, but may not correspond exactly to the actual average length of time between visits. The enrolment visit (visit 1, month 0) involved administration of a questionnaire (to collect data on socio-demographics and the participants' medical, gynaecological and sexual history), a clinical exam (including a speculum exam and visual inspection with acetic acid and lugol's iodine, VIA/VILI, to identify cervical lesions), collection of cervico-vaginal specimens (to test for infections and to determine vaginal cytology) and a blood sample. Participants were counselled and treated for genital infections as per local guidelines if any abnormalities were found. A second visit (visit 2) was scheduled shortly after visit 1 (allowing time for any results from visit 1 to become available) at which colposcopy and cervical biopsies (if indicated, see section 4.2.6) were performed. The result of the biopsy was communicated confidentially to participants and treatment provided (if required) at a further visit (visit 2b).

Participants were asked to return at month 6 (visit 3) and month 12 (visit 4) for routine monitoring of their HIV infection and administration of a short questionnaire. Additionally, a cervical (for HPV testing) and blood (for CD4 count) sample was taken at visit 4.

At month 18, participants were recalled for re-examination (visit 5 and 6/6a). Procedures and participant management at these visits were similar to visit 1 and

2/2b. A pregnancy check was carried out at visit 5/6, since this would be a contraindication for cervical biopsy. The VMB-HARP sub-study utilised vaginal samples collected at visits 1 and 5.

If women were not already on ART, it was initiated during the study once the measured CD4 count was 350 cells/mm^3 or below, which was in accordance with the WHO guidelines that were in place during HARP study implementation (WHO 2010).

4.2.3 Testing for bacterial vaginosis and candidiasis

Vaginal smears were taken at visit 1 (month 0) and evaluated for the presence of yeasts (referred to as candidiasis) and bacterial vaginosis (BV) by the Nugent's score method where BV was defined as a score of 7-10. For the South African samples, this testing was carried out by the National Health Laboratory Service STI Reference Centre in Johannesburg, which subscribes to an international external quality assurance programme.

4.2.4 Cervical cytology

Cervical cytology was assessed by examination of cervical brush Pap smears taken at visit 1 (month 0) and visit 5 (month 18) and classified according to the Bethesda classification system (Solomon et al 2002). This assessment was performed locally by two different pathologists/cytologists and assessed by a third observer in case of discordance. Quality control was carried out at the University of Montpellier 1 (UM1) pathology department on 5-10% of slides.

4.2.5 Blood sampling and STI testing

Blood samples for measurement of HIV plasma viral load (PVL; COBAS Taqman, Roche Diagnostics, Johannesburg, South Africa; lower limit of detection of 40 copies/ml), and for syphilis serology (combined *Treponema pallidum* haemagglutination and rapid plasma reagin using Immutrep carbon antigen RPR, Omega Diagnostics, Cape Town) and herpes simplex virus-2 serology (Kalon IgG2 ELISA, Kalon Diagnostics, Guildford, UK) were taken at visit 1 (month 0) and tested locally. Additionally, CD4+ lymphocyte counts (FACScount, Becton-Dickinson, Franklin Lakes, New Jersey, USA) were obtained at visit 1 (month 0), visit 3 (month 6), visit 4 (month 12), and visit 5 (month 18) as part of routine HIV monitoring. Cervical swabs to test for infection with *Neisseria gonorrhoeae*, *Chlamydia*

trachomatis, *Mycoplasma genitalium*, and *Trichomonas vaginalis* by PCR (APTIMA Combo, Gen-Probe, San Diego, California, USA) were taken at visit 1 (month 0).

4.2.6 Outcome measures: HPV and CIN status

Pap smears were collected at visits 1 and 5 (month 0 and 18) to determine HPV status by genotyping, which was carried out at UM-1 using the InnoLipA HPV genotyping Extra Assay (Innogenetics, Courtaboeuf, France). This assay amplifies and detects HPV DNA and allows the identification of all high-risk and probable high-risk HPV types (16, 18, 26, 31, 33, 35, 39, 45, 51, 52, 53, 56, 58, 59, 66, 68, 73, 82) as well as a number of low-risk types (6, 11, 40, 43, 44, 54, 70) and additional types (69, 71, 74). For HPV types that cannot be genotyped in this way, sequencing was carried out. UM-1 subscribes to an annual European external quality assurance scheme for HPV testing, and the results were 100% satisfactory.

CIN status was determined according to the following. VIA/VILI was carried out by a trained nurse midwife at visits 1 and 5 (months 0 and 18). Colposcopy was carried out by a trained gynaecologist at visits 2 and 6 (scheduled once results from visit 1 and 5 were available, respectively) and four-quadrant cervical biopsies (and directed biopsies in case of lesions) were taken if any one of the following results were obtained: positive HR-HPV test (by either genotyping or qualitative DNA test: Digene HC2 test at visit 1 and CareHPV test at visit 5), presence of cytological abnormalities (any grade above ASC-US), abnormal VIA/VILI results, or abnormalities detected on colposcopy. Histology was assessed by local pathology laboratories according to the CIN classification system (Richart classification). A consensus pathology review (endpoint committee, EPC) was completed to review the histology results for all samples classified locally as positive for cervical precancer (histological grade of CIN2+) and for 5-10% of negative samples at the end of the two phases of the research. This involved pathologists from the study sites, UM-1 and an external expert from the Institut Catala d'Oncologia, Barcelona. The results of the EPC will be used for the VMB-HARP sub-study. For convenience and in order to match the corresponding HPV results, the biopsy results from visit 2 and 6 are referred to as visit 1 and 5 in the text that follows, respectively. Since CIN status at endline in women who were coinfected with HIV and HPV at baseline was the main outcome of interest in the HARP study, a relatively long sampling interval of 18 months was chosen to allow time for the development of dysplasia.

4.3 Methods: VMB-HARP Study

4.3.1 Sample characteristics

Samples used for the VMB-HARP study consisted of vaginal swabs stored in 2 ml BoonFix® (a fixative containing ethanol and polyethylene glycol) at room temperature until DNA extraction. These were collected at visits 1 and 5 of the HARP study by swabbing the posterior fornices and lateral walls of the vagina using a Dacron swab during speculum examination. The swab was then immediately placed into BoonFix® medium. Ethical approval for the determination of the vaginal VMB from these samples had been obtained from the local ethics committees at Wits University in Johannesburg, South Africa and the London School of Hygiene and Tropical Medicine, UK; specific approval for VMB-HARP was obtained from the University of Liverpool, UK (Physical Interventions Sub-Committee) and by submission of an amendment to the VMB-HARP protocol from the ethics committee at Wits University.

4.3.2 Sub-sampling for VMB-HARP

The VMB-HARP study has two objectives: 1) to determine the association between the VMB type and incidence, persistence and clearance of HR-HPV in HIV-infected South African women and 2) to determine the association between the VMB type and the presence of CIN2+ lesions (prevalent or incident) among HIV-infected African women. In order to study these two aims, samples were split into two groups. Women with ≤CIN1 at both visits comprise part 1 of the study and data obtained in this part of the study will be used to answer objective 1. Women with CIN2+ at either visit comprise part 2 of the study and data obtained in this part of the study will be used to answer objective 2. Women who had persistent HR-HPV infection and were analysed in part 1 of the study will also be used as the control group for part 2 of the study. Hence, the case and control definitions are as follows:

Part 1:

- Persistent HR-HPV: At least one identical HR-HPV type present at visit 1 and visit 5.
- Incident HR-HPV: No HR-HPV present at visit 1 and at least one HR-HPV type present at visit 5.
- Cleared HR-HPV: At least one HR-HPV type present at visit 1 and no HR-HPV present at visit 5.
- Type swap: At least one HR-HPV type present at visit 1 and visit 5, but none of the types are identical between visits.
- Controls: Negative for HR-HPV at both time points.

Part 2:

- Incident CIN2+: <CIN2 present at visit 1 and CIN2+ present at visit 5 (for prospective analysis).
- Persistent CIN2+: CIN2+ at visit 1 and visit 5 (for cross-sectional analysis).
- Cleared CIN2+: CIN2+ present at visit 1 and <CIN2 present at visit 5.
- Prevalent CIN2+: CIN2+ at visit 1 or visit 5 with data missing for the other visit (for cross-sectional analysis).
- Controls: Persistent HR-HPV at both time points with <CIN2 (same as persistent HR-HPV in part 1 above).

All samples from South African women who had a valid HPV genotyping and CIN result (either a valid biopsy result, or biopsy was not indicated) for visits 1 and 5, and for whom BoonFix® samples were available for both visits, were analysed. Additionally, since we were particularly interested in the VMB of women with CIN2+, we also analysed samples from women who had CIN2+ at the corresponding visit, even if their data were missing and/or the sample from the other visit was missing (see Figure 4.2).

4.3.3 DNA extraction

DNA extraction was carried out according the methods developed in Chapter 2, as follows. Samples were thoroughly mixed by vortexing. Then the swab head and 100 µl of liquid were subjected to 30 min of lysis at 37°C using enzymatic lysis buffer containing lysozyme from chicken egg white (the recommended pretreatment for Gram-positive bacteria as per the Qiagen DNeasy Blood and Tissue kit Handbook). Proteinase K and Buffer AL (Qiagen) were added and incubated at 56°C for 30 min, followed by the remaining steps in the kit's spin column protocol, in accordance with the manufacturer's instructions. DNA was eluted in 75 µl of elution buffer. The genomic DNA concentration was measured with the Qubit Fluorometer using the dsDNA HS Assay kit. A negative extraction control (containing 100 µl nuclease free water) was included in each extraction run.

4.3.4 Amplicon library preparation and DNA sequencing

The V3-V4 region of the 16S rRNA gene contained in 10 µl of DNA extract was amplified in a 25 µl reaction as described previously (see section 3.2.2). In order to optimise PCR product yield – according to the results of a pilot run – sample DNA extracts ≤ 25 ng/µl were not diluted, extracts > 25 ng/µl and ≤ 100 ng/µl were diluted 1:4, and extracts > 100 ng/µl were diluted 1:8 prior to PCR. Each PCR run included positive controls (as described in Chapter 3) and a negative PCR control (containing

nuclease free water). PCR products were purified, eluted in a volume of 10 µl TE buffer (Sigma-Aldrich) and quantified using the Qubit Fluorometer with the dsDNA HS Assay kit to determine amplicon yield. Purified PCR amplicons measuring ≥ 2 ng/µl were run on a 2% agarose gel at 100V to verify purity of the amplicon. While all amplicons were sequenced, those measuring ≤ 0.2 ng/µl additionally underwent repeated PCR and sequencing to reduce the risk of having to exclude samples due to insufficient read counts. Pooled amplicons were sequenced at the University of Liverpool Centre for Genomics Research on the Illumina HiSeq platform (2x300bp; Illumina) on two separate runs, one consisting of two lanes on the same flowcell (designated "X" and "Y") and one consisting of a single lane (designated "Z"). These are the same sequencing runs as described in Chapter 3. Negative extraction controls, negative PCR controls and the positive controls described in Chapter 3 were also sequenced on these sequencing runs.

4.3.5 Bioinformatics

Sequencing data obtained was processed according to the Swarm pipeline detailed in section 3.2.3. Additionally, species assignments were made using DADA2's *assignSpecies* function against the Silva v. 128 database, allowing multiple matches (option *allowMultiple* enabled). Bacterial vaginosis-associated bacteria (BVAB) BVAB1, BVAB2, *Mageeibacillus indolicus* (BVAB3) and BVAB TM7, and *Fenollaria massiliensis* are not contained in this database and were identified manually from representative DNA sequences. Operational taxonomic unit (OTU) representative sequences (i.e. most abundant in OTU) that made up at least 0.05% of the total read count and had no species assignment were additionally BLAST searched (Altschul et al 1990) to identify any identical matches in the NCBI database. If these were found, OTU descriptors were left unchanged unless they contradicted the BLAST result. In this case, the next concurring higher level descriptor was assigned. Note that this pipeline may result in multiple OTUs for a single species (due to different strains of the same species having differences in their DNA sequence in the V3-V4 region of the 16S rRNA gene) and may result in a single OTU that matches more than one species (if there are multiple sequences in the Silva v. 128 database that have exactly the same DNA sequence in the V3-V4 region and therefore cannot be differentiated).

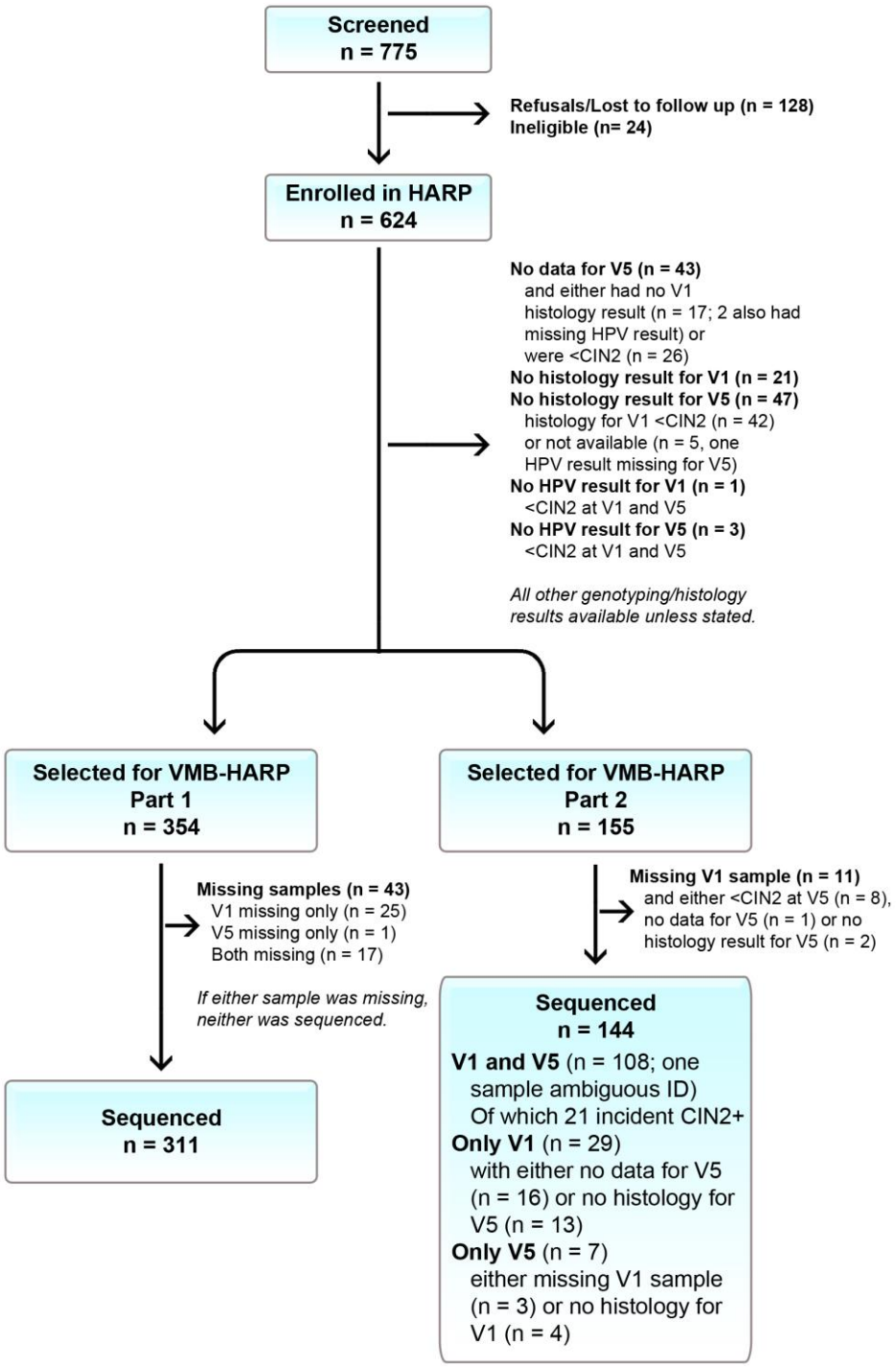


Figure 4.2 Sample selection for the VMB-HARP sub-study. The numbers given represent study participants, rather than individual samples (of which each patient may have up to 2). Of the 43 women that were excluded because of missing data for visit 5, 19 were lost to follow up, 8 moved away, 7 were withdrawn by the clinician (2 of which were referred for suspected cervical carcinoma without a biopsy result), 5 withdrew due to personal circumstances and/or being unwilling to undergo further testing and 4 died (one of which died from probable HIV-related causes, the other causes of death were unknown). The woman who was excluded from part 2 due to having a missing sample for visit 1 and no data for visit 5 had been withdrawn due to hysterectomy. Of the 16 women that had only the visit 1 sample sequenced for part 2 due to missing data for visit 5, 2 were lost to follow up, 8 moved away, 3 were withdrawn by the clinician (2 of which were withdrawn because of hysterectomy), 2 withdrew due to personal reasons and one died due to HIV-related causes.

4.3.6 OTU table generation

The raw counts obtained from the above bioinformatics pipeline were processed as follows. Singletons were removed. In a minority of reads (<0.5%), there was incomplete removal of the barcode due to sequencing errors in this region. This resulted in these longer reads being assigned to a different OTU (rather than the OTU that they would have been assigned to without the incompletely removed barcode). Where the reference sequences to these OTUs were an exact match (i.e. when aligned, the sequences were identical in the region of overlap), they were merged in R version 3.2.2 (R Core Team 2015) using an algorithm written for this purpose. Following this, any OTUs with a read count under 100 reads were removed. OTUs were removed as likely contaminants if they were present in more than one negative extraction control or more than one negative PCR control, had a higher percentage in at least one negative control compared to all other samples and if they had an absolute read count that was not significantly higher by t test in samples compared with negative controls (to avoid removing OTUs that are present in negative controls due to barcode switching). Any OTUs identified as human or chloroplast DNA were removed. For any samples that had been sequenced twice, the profile with the lowest read count was removed. Based on rarefaction curves and the read counts of repeat samples where concurrence was poor (see Appendix F), a minimum read count of 1000 reads was decided upon. Samples were therefore rarefied to 1039 reads (i.e. the smallest read count ≥1000 reads) using the GUniFrac package version 1.0 in R version 3.4.1. Rarefied counts were used for all downstream analyses unless otherwise stated.

4.3.7 Data analysis

Calculation of alpha and beta diversity measures, clustering of VMB profiles, generation of graphs, nonmetric multidimensional scaling and statistical analyses were performed in R version 3.4.1 (R Core Team 2015) and using the vegan package version 2.4-2 (Oksanen et al 2015). Multinomial regression modelling additionally made use of the nnet package version 7.3.12 (Venables and Ripley 2013). NMDS results were plotted using the plotly package 4.7.1 (Sievert et al 2017). VMB profiles were clustered using complete linkage hierarchical clustering on the Euclidean distance metric on rarefied read counts, using a cut-off that visually resulted in a high degree of similarity of samples within a cluster (cut-off = 650). These fine-scale clusters were then manually grouped into VMB types, based on pre-existing knowledge of the prevalence of and clinical/biological properties of

these compositions. See section 4.4.4 for further details. All VMB profiles from both visits were used for clustering and overall VMB profile descriptions (not including repeat sample profiles that were discarded and those with read counts <1000, see section 4.4.2).

In order to determine if there was an association between different outcome categories and individual OTUs, we used the Linear Discriminant Analysis effect size (LEfSe) algorithm (Segata et al 2011). The significance level for the Kruskal-Wallis test was 0.05 and the threshold on the logarithmic Linear Discriminant Analysis score was 1.5.

Differences in characteristics between study groups were compared for categorical variables using Pearson's chi-squared with Yates' continuity correction or Fisher's exact test (where there were expected values below 5). For large contingency tables with more than two rows and columns the P value was estimated with the Monte Carlo method (50,000 simulations). Continuous variables were compared using the Kruskal Wallis test. Any significant results comparing the 5 HR-HPV or 5 CIN categories were followed up with appropriate post-hoc pairwise tests with Holm-Bonferroni correction (chi-squared or Fisher's exact test using R package fifer v 1.1 for categorical variables and Mann–Whitney U test for continuous variables). Differences in *Lactobacillus* relative abundance and alpha diversity were determined using the Mann–Whitney U test with Holm-Bonferroni correction. In addition, beta (between-sample) diversity using Bray Curtis similarity (ranges from 0 to 100%) was calculated between the two samples taken from each woman at visits 1 and 5 and the Kruskal Wallis test was used to assess changes over time.

Epidemiological analyses used two different outcomes: the 5 HR-HPV categories (part 1) and the 5 CIN categories (part 2) described in section 4.3.2. The VMB composition was always the main predictor (forced into all models), and the following VMB composition measures were used in different models: VMB types by hierarchical clustering (N=7 VMB community types), *Lactobacillus* genus relative abundance (a proportion between 0-100%), and alpha (within-sample) diversity using the Simpson index (1-D; ranges from 0 to 1). Univariable multinomial logistic regression models were used to determine the unadjusted association between the VMB variables and the outcome groups. Multivariable multinomial logistic regression models were used to adjust for confounding. Potential confounders were

selected *a priori*, based on their potential association with both the VMB and HR-HPV infection and/or CIN. These were included in the model by forward selection if they were found to be associated with the outcome in univariable regression at a significance level of <0.1, and if the model did not already include a predictor with which they were highly correlated.

4.4 Results

4.4.1 VMB-HARP study population characteristics

Women selected for the VMB-HARP study who had a valid sequencing result described their ethnic group as either black (n=449) or coloured (n=1). The median age of these women was 34 years at enrolment, ranging from 25-50 and the median time between visit 1 and visit 5 was 485 days (15.9 months). Some statistically significant differences were identified between women selected for VMB-HARP and those that were excluded (Table 4.1). A significantly higher proportion of women selected for VMB-HARP was on HIV ART at study commencement (P <0.001) and they had a significantly lower median plasma viral load (P = 0.043).

A number of significant differences were also found between the VMB-HARP study groups described in section 4.3.2 (Table 4.1). By design, differences in HR-HPV and any HPV infection were highly significant among the 5 HR-HPV groups in part 1 of the study (P <0.001), with post hoc pairwise tests finding differences occurring as expected between groups. There was also a significant difference in HR-HPV at baseline among the 5 CIN groups in part 2 of the study. However, post-hoc pairwise testing identified a statistically significant difference only between the group with persistent HR-HPV but no CIN2+ and the cleared CIN2+ groups which can also be explained by design. Additionally, post-hoc pairwise tests found that women that cleared CIN2+ were significantly 1) more likely to be using injectable contraceptives compared to women who had either persistent HR-HPV infection without CIN2+ (controls; P = 0.002) and those that had incident CIN2+ (P = 0.021) and 2) more likely to be using any hormonal contraceptives compared to controls (P = 0.028). All other post-hoc pairwise tests did not reach statistical significance after Holm–Bonferroni correction.

Table 4.1 Characteristics of women selected for the VMB-HARP study by study group (continued on next page). Numbers of women are shown together with the percentage of women with this characteristic in the study group (for categorical variables) or median value and interquartile range (for continuous variables). Kruskal Wallis was used for continuous variables and Chi^2 or Fisher's exact test were used for categorical variables, except for large contingency tables with more than two rows and columns where the p value was estimated using the Monte Carlo method (50,000 simulations).

VMB-HARP STUDY PART	**PART 1**				**1&2**	**PART 2**								
	Negative (controls)	**Incident HR-HPV**	**Cleared HR-HPV**	**Type swap HR-HPV**	**Persistent HR-HPV**	**Incident CIN2+**	**Cleared CIN2+**	**Persistent CIN2+**	**Prevalent CIN2+**	**Part 1**	**Part 2**	**VMB-HARP (all)**	**HARP (all)**	
	(n = 38)	(n = 43)	(n = 68)	(n = 69)	(n = 93)	(n = 22)	(n = 64)	(n = 25)	(n = 33)	p value[1]	p value[2]	(n = 455)	(n = 624)	p value[3]
Median age at baseline (IQR)	38 (34-42)	36 (33-41)	36 (31-39)	34 (30-40)	33 (30-39)	34 (30-35)	34 (30-36)	33 (30-36)	32 (26-38)	0.051	0.936	34 (30-39)	34 (30-39)	0.377
Self-reported current smoker	1 (2.6%)	0 (0%)	4 (5.9%)	7 (10.1%)	6 (6.5%)	1 (4.5%)	3 (4.7%)	1 (4.0%)	2 (6.1%)	0.213	0.989	25 (5.5%)	34 (5.4%)	1.000
SEXUAL BEHAVIOUR AND CONTRACEPTION AT BASELINE														
Ever used any form of contraception	36 (94.7%)	42 (97.7%)	67 (98.5%)	65 (94.2%)	89 (95.7%)	22 (100%)	62 (96.8%)	24 (96.0%)	32 (97.0%)	0.697	1.000	439 (96.5%)	601 (96.3%)	0.897
Currently using any hormonal contraceptive	10 (26.3%)	8 (18.6%)	15(22.1%)	12 (17.4%)	19 (20.4%)	5 (22.7%)	27 (42.2%)	12 (48.0%)	7 (21.2%)	0.846	0.006*	115 (25.3%)	152 (24.4%)	0.442
Currently using oral contraceptive or patch	1 (2.6%)	0 (0%)	3 (4.4%)	1 (1.4%)	8 (8.6%)	4 (18.2%)	2 (3.1%)	2 (8.0%)	2 (6.1%)	0.120	0.216	23 (5.0%)	29 (4.6%)	0.562
Currently using injectable contraceptive	9 (23.7%)	7 (16.3%)	11 (16.2%)	8 (11.6%)	11 (11.8%)	1 (4.5%)	25 (39.1%)	9 (36.0%)	5 (15.2%)	0.438	<0.001*	86 (18.9%)	116 (18.6%)	0.832
Currently using contraceptive implants	0 (0%)	1 (2.3%)	1 (1.5%)	3 (4.3%)	0 (0%)	0 (0%)	0 (0%)	1 (4.0%)	0 (0%)	0.187	0.198	6 (1.3%)	7 (1.1%)	0.735
Condom use in last 3 months														
No recent sex	9 (23.7%)	8 (18.6%)	13 (19.1%)	9 (13.0%)	20 (21.5%)	1 (4.5%)	7 (10.9%)	6 (24.0%)	8 (24.2%)	0.437	0.094	81 (17.8%)	117 (18.8%)	0.680
never	2 (5.3%)	3 (7.0%)	3 (4.4%)	7 (10.1%)	0 (0%)	3 (13.6%)	2 (3.1%)	2 (8.0%)	2 (6.1%)			24 (5.3%)	31 (5.0%)	
sometimes	10 (26.3%)	9 (20.9%)	17 (25.0%)	18 (26.1%)	27 (29.0%)	7 (31.8%)	23 (35.9%)	8 (32.0%)	10 (30.3%)			129 (28.4%)	172 (27.6%)	
always	17 (44.7%)	23 (53.5%)	35 (51.5%)	35 (50.7%)	46 (49.5%)	11 (50.0%)	32 (50.0%)	9 (36.0%)	13 (39.4%)			221 (48.6%)	304 (48.7%)	
Have a current regular male sexual partner	29 (76.3%)	34 (79.1%)	53 (77.9%)	59 (85.5%)	74 (79.6%)	21 (95.5%)	55 (85.9%)	19 (76.0%)	25 (75.8)	0.763	0.244	369 (81.1)	502 (80.4%)	0.577

[1]P value for differences between VMB-HARP study groups in part 1 (negative, incident, cleared, type-swap and persistent HR-HPV).
[2]P value for differences between VMB-HARP study groups in part 2 (persistent HR-HPV and incident, cleared, persistent and prevalent CIN2+).
[3]P value for differences between women selected for VMB-HARP and those in the HARP study not selected for VMB-HARP.

Table 4.1 (cont.)

VMB-HARP STUDY PART	PART 1				1&2	PART 2								
	Negative (controls)	Incident HR-HPV	Cleared HR-HPV	Type swap HR-HPV	Persistent HR-HPV	Incident CIN2+	Cleared CIN2+	Persistent CIN2+	Prevalent CIN2+	Part 1	Part 2	VMB-HARP (all)	HARP (all)	
Number of lifetime male sexual partners														
1	2 (5.3%)	2 (4.7%)	4 (5.9%)	0 (0%)	1 (1.1%)	0 (0%)	2 (3.1%)	0 (0%)	0 (0%)	0.108	0.077	11 (2.4%)	17 (2.7%)	0.789
2-4	21 (55.3%)	21 (48.8%)	21 (30.9%)	27 (39.1%)	47 (50.5%)	12 (54.5%)	34 (53.1%)	12 (48.0%)	18 (54.5%)			213 (46.8%)	284 (45.5%)	
5-9	6 (15.8%)	13 (30.2%)	23 (33.8%)	21 (30.4%)	31 (33.3%)	2 (9.1%)	15 (23.4%)	10 (40.0%)	8 (24.2%)			129 (28.4%)	178 (2z8.5%)	
10+	3 (7.9%)	4 (9.3%)	9 (13.2%)	7 (10.1%)	7 (7.5%)	0 (0%)	5 (7.8%)	0 (0%)	1 (3.0%)			36 (7.9%)	51 (8.2%)	
unknown	6 (15.8%)	3 (7.0%)	11 (16.2%)	14 (20.3%)	7 (7.5%)	8 (36.4%)	8 (12.5%)	3 (12.0%)	6 (18.2%)			66 (14.5%)	94 (15.1%)	
Number of male sexual partners in last 3 months														
0	9 (23.7%)	7 (16.3%)	12 (17.6%)	9 (13.0%)	19 (20.4%)	1 (4.5%)	6 (9.4%)	6 (24.0%)	8 (24.2%)	0.305	0.174	77 (16.9%)	111 (17.8)	0.434
1	29 (76.3%)	34 (79.1%)	50 (73.5%)	51 (73.9%)	72 (77.4%)	19 (86.4%)	54 (84.4%)	19 (76.0%)	24 (72.7%)			352 (77.4%)	476 (76.3)	
2-4	0 (0%)	1 (2.3%)	5 (7.4%)	7 (10.1%)	1 (1.1%)	1 (4.5%)	3 (4.7%)	0 (0%)	1 (3.0%)			19 (4.2%)	29 (4.6%)	
5-9	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (4.5%)	0 (0%)	0 (0%)	0 (0%)			1 (0.2%)	1 (0.2%)	
10+	0 (0%)	1 (2.3%)	1 (1.5%)	1 (1.4%)	1 (1.1%)	0 (0%)	1 (1.6%)	0 (0%)	0 (0%)			5 (1.1%)	5 (0.8%)	
unknown	0 (0%)	0 (0%)	0 (0%)	1 (1.4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)			1 (0.2%)	2 (0.3%)	
Median age at first intercourse (IQR)	18 (16-19)	17 (16-18)	17 (16-19)	17 (16-19)	18 (17-19)	17 (16-19)	18 (17-19)	18 (16-19)	18 (17-19)	0.061	0.623	18 (16-19)	18 (16-19)	0.801
Practice vaginal cleansing at least weekly	17 (44.7%)	17 (39.5%)	30 (44.1%)	26 (37.7%)	32 (34.4%)	10 (45.5%)	20 (31.3%)	13 (52.0%)	14 (42.4%)	0.710	0.328	179 (39.3%)	254 (40.7%)	0.295
Ever earned money/drugs/food/bed for sex	2 (5.3%)	1 (2.3%)	7 (10.3%)	11 (15.9%)	7 (7.5%)	0 (0%)	5 (7.8%)	2 (8.0%)	0 (0%)	0.140	0.339	35 (7.7%)	42 (6.7%)	0.164

[1]P value for differences between VMB-HARP study groups in part 1 (negative, incident, cleared, type-swap and persistent HR-HPV).
[2]P value for differences between VMB-HARP study groups in part 2 (persistent HR-HPV and incident, cleared, persistent and prevalent CIN2+).
[3]P value for differences between women selected for VMB-HARP and those in the HARP study not selected for VMB-HARP.

Table 4.1 (cont.)

VMB-HARP STUDY PART	PART 1				1&2	PART 2								
	Negative (controls)	Incident HR-HPV	Cleared HR-HPV	Type swap HR-HPV	Persistent HR-HPV	Incident CIN2+	Cleared CIN2+	Persistent CIN2+	Prevalent CIN2+	Part 1	Part 2	VMB-HARP (all)	HARP (all)	
HIV-RELATED FACTORS														
Median CD4 count cells/mm^3 at baseline	457 (347-595)	439 (318-562)	483 (348-604)	403 (281-523)	433 (335-555)	438 (355-547)	398 (298-580)	377 (245-436)	326 (214-514)	0.207	0.284	425 (318-566)	428 (322-581)	0.149
Median baseline plasma viral load (HIV-1 copies per millilitre log10)	2.6 (2.0-3.7)	2.7 (1.9-3.5)	2.2 (1.6-2.8)	2.4 (1.6-3.4)	2.8 (1.6-3.9)	3.2 (2.1-4.3)	3.0 (1.9-4.5)	3.9 (1.7-4.6)	2.9 (1.6-4.0)	0.102	0.602	2.6 (1.6-3.9)	2.7 (1.6-4.0)	0.043*
Antiretroviral therapy														
not on ART throughout study	10 (26.3%)	7 (16.3%)	14 (20.6%)	14 (20.3%)	26 (28.0%)	10 (45.5%)	17 (26.2%)	8 (32.0%)	6 (18.2%)	0.142	0.164	112 (24.6%)	181 (29.0%)	<0.001*
on ART at study commencement	25 (65.8%)	35 (81.4%)	54 (79.4%)	54 (78.3%)	61 (65.6%)	11 (50.0%)	38 (58.5%)	15 (60.0%)	26 (78.8%)			319 (70.0%)	406 (65.1%)	
started ART during study	3 (7.9%)	1 (2.3%)	0 (0%)	1 (1.4%)	6 (6.5%)	1 (4.5%)	10 (15.4%)	2 (8.0%)	1 (3.0%)			25 (5.5%)	37 (5.9%)	
VAGINAL INFECTIONS AT BASELINE														
Bacterial vaginosis (Nugent score 7-10)	14 (37.8%)	18 (41.9%)	23 (33.8%)	37 (55.2%)	41 (46.1%)	10 (45.5%)	25 (39.7%)	11 (45.8%)	13 (40.6%)	0.131	0.938	192 (43.1%)	254 (41.5%)	0.210
Candidiasis	2 (5.4%)	4 (9.3%)	2 (2.9%)	5 (7.5%)	9 (9.9%)	2 (9.1%)	6 (9.5%)	1 (4.0%)	3 (9.4%)	0.474	0.973	34 (8.1%)	53 (8.7%)	0.193
SEXUALLY TRANSMITTED INFECTIONS AT BASELINE														
Any HPV (genotyping)	22 (57.9%)	19 (44.2%)	68 (100%)	69 (100%)	93 (100%)	22 (100%)	60 (93.8%)	24 (96.0%)	33 (100%)	<0.001*	0.058	410 (90.1%)	552 (88.7%)	0.711
HR-HPV (genotyping)	0 (0%)	0 (0%)	68 (100%)	69 (100%)	93 (100%)	21 (95.5%)	58 (90.6%)	24 (96.0%)	30 (90.9%)	<0.001*	0.011*	363 (79.8%)	491 (78.9%)	0.823
Chlamydia trachomatis	1 (2.6%)	5 (11.6%)	1 (1.5%)	3 (4.3%)	4 (4.3%)	0 (0%)	5 (7.8%)	1 (4.0%)	4 (12.1%)	0.201	0.360	24 (5.3%)	31 (5.0%)	0.741
Neisseria gonorrhoeae	0 (0%)	2 (4.7%)	1 (1.5%)	2 (3.0%)	1 (1.1%)	1 (4.5%)	0 (0%)	1 (4.0%)	2 (6.1%)	0.562	0.103	10 (2.2%)	13 (2.1%)	1.000
HSV-2	35 (92.1%)	41 (95.3%)	67 (98.5%)	68 (98.6%)	88 (95.7%)	21 (95.5%)	60 (93.8%)	21 (87.5%)	31 (93.9%)	0.361	0.623	432 (95.4%)	591 (95.2%)	1.000
Mycoplasma genitalium	4 (10.5%)	3 (7.0%)	4 (5.9%)	7 (10.1%)	6 (6.5%)	4 (18.2%)	4 (6.3%)	4 (16.0%)	2 (6.1%)	0.809	0.219	38 (8.4%)	46 (7.4%)	0.211
Active syphilis (*Treponema pallidum*)	0 (0%)	0 (0%)	1 (1.5%)	0 (0%)	0 (0%)	2 (9.1%)	0 (0%)	0 (0%)	0 (0%)	0.479	0.008*	3 (0.7%)	7 (1.1%)	0.094
Trichomonas vaginalis	3 (7.9%)	8 (18.6%)	9 (13.2%)	17 (24.6%)	12 (12.9%)	3 (13.6%)	11 (17.2%)	4 (16.0%)	4 (12.1%)	0.134	0.939	71 (15.6%)	101 (16.2%)	0.651

[1]P value for differences between VMB-HARP study groups in part 1 (negative, incident, cleared, type-swap and persistent HR-HPV).
[2]P value for differences between VMB-HARP study groups in part 2 (persistent HR-HPV and incident, cleared, persistent and prevalent CIN2+).
[3]P value for differences between women selected for VMB-HARP and those in the HARP study not selected for VMB-HARP.

4.4.2 Sequencing results

A total of 125,034,772 paired 16S rRNA sequence reads (V3-V4 region) were generated from the VMB-HARP samples (946 microbiome profiles, of which 71 were experimental (PCR) replicates, were generated from 875 samples from 455 women). After error correction (0.2% of total reads discarded), paired-end alignment (3.1% discarded) and removal of sequences containing ambiguous bases (0.1% discarded), singletons (4.8% discarded), small OTUs under 100 reads (0.3% discarded), chimeric sequences (1.0% discarded) and contaminants (0.5% discarded), 112,558,944 sequences remained. After discarding contaminant OTUs and those with a total read count below 100, 1983 OTUs remained. After discarding samples with a read count <1000 reads (resulting in loss of four samples from four different women, two of whom were in part 1 of VMB-HARP and two in part 2), the median read count was 122,490 reads, ranging from 1039 to 859,842 reads (not including repeat sample profiles that were discarded). In one case there were two samples for visit 5 for the same subject, which produced distinct VMB profiles. These profiles were used in the clustering of samples, but were removed from downstream analyses. The presence of two samples for the same subject may have been due to sample mislabelling and since the identity of neither sample could be ascertained with complete certainty, neither profile was used in the final analysis.

Negative PCR controls (12 samples) contained a median of only 208 raw reads, despite a similar or higher average volume of amplicon added to the final amplicon pool when compared to samples. Most of these reads failed to align (82%). Using the criteria described in the methods section, only five OTUs were identified as PCR contaminants (see Appendix G), which amounted to <0.01% of total classified sample reads. PCR contamination was therefore considered negligible. Contamination originating from the extraction process was higher, with 111 contaminant OTUs identified (see Appendix G) making up 0.5% of total classified sample reads. By far the most common contaminant in all negative extraction controls was *Rhodanobacter glycinis*/*terrae* with a relative abundance of 43-96%. This OTU was more than 450 times as abundant in the negative extraction controls when compared to the most common PCR contaminant in negative PCR controls (*Achromobacter denitrificans*/*ruhlandii*/*xylosoxidans*), indicating that contamination from the DNA extraction process was much higher than that occurring during PCR. This is supported by the observation that the *Rhodanobacter* OTU was more prevalent in vaginal samples compared to the *Achromobacter* OTU (median relative

abundance in samples of *Rhodanobacter* was 0.020%, while *Achromobacter* was absent from the majority of samples). Of the total classified reads in negative PCR and negative extraction controls 30.3% and 97.6% were identified as contaminants, respectively. The majority of remaining reads belonged to the top 10 most common OTUs in terms of total read count in the entire dataset (56.0% and 69.5% of non-contaminant reads in the negative PCR and negative extraction controls, respectively). This is consistent with barcode switching, although cross-contamination from samples cannot be ruled out.

4.4.3 Vaginal microbiome composition in the VMB-HARP population

The OTUs with the highest read counts in the VMB-HARP population as a whole were *Lactobacillus iners* (36.5% of classified reads), *Gardnerella vaginalis* OTU 0 (9.2%), *Lactobacillus acidophilus*/*casei*/*crispatus*/*gallinarum* (7.9%), *G. vaginalis* OTU 1 (4.5%), BVAB1 (4.5%), *Sneathia amnii*/*sanguinegens* (4.0%), *Atopobium vaginae* (2.5%), *Megasphaera* OTU 0 (2.5%), *Prevotella* OTU 0 (1.8%), *G. vaginalis* OTU 2 (1.7%), *Megasphaera* OTU 1 (1.6%), *Dialister* (1.5%), Coriobacteriales (1.5%), *Lactobacillus jensenii* (1.3%), *Sneathia sanguinegens* (1.3%), *Prevotella bivia* (1.0%) and *Streptococcus agalactiae*/*pyogenes* (1.0%). All other OTUs made up less than 1% of classified reads. The distribution of these OTUs across samples from both visits is shown in Figure 4.3.

The use of Swarm for clustering of reads into OTUs allowed for higher resolution compared with older methods based on an arbitrary global similarity threshold (such as USEARCH) and allowed distinction between closely related sequences. This was particularly true in the case of *G. vaginalis*, where a total of 4 OTUs were identified, sharing between 98.8-99.8% sequence similarity. *G. vaginalis* OTU 0 shares 100% sequence identity with the published complete genome sequences of *G. vaginalis* strains 409-05 and GV37. The former was isolated from the vagina of an asymptomatic woman with a Nugent score of 9 (Yeoman et al 2010), while the latter was isolated from the blood of a woman with toxic encephalopathy (Tankovic et al 2017). *G. vaginalis* OTU 1 shares 100% sequence identity with three complete genomes, those of *G. vaginalis* strains HMP9231, ATCC14018 (=JCM10026) and ATCC 14019. HMP9231 was isolated from endometrium, while the other two were isolated from women with symptomatic BV (Cornejo et al 2018, Yeoman et al 2010). There are no complete genome sequences available with an identical 16S rRNA sequence to *G. vaginalis* OTUs 2 and 3, although identical 16S rRNA gene

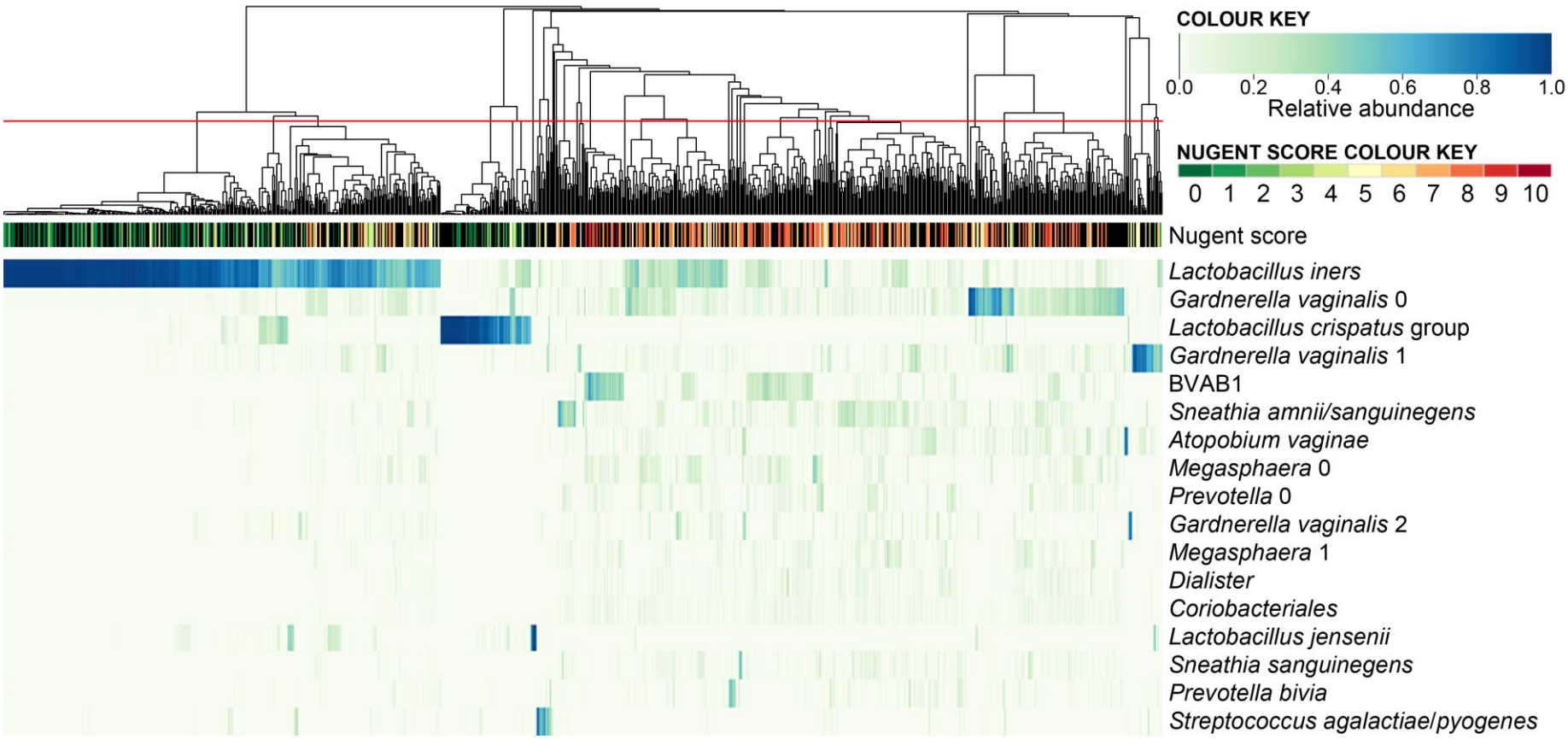


Figure 4.3 Heat map showing the relative abundance of all OTUs that made up at least 1% of reads in all VMB-HARP samples. The tree shown represents the results of complete linkage hierarchical clustering on the Euclidean distance. The cut-off for separating sequences into fine scale clusters is indicated by a red line. For samples from visit 1, the top bar indicates the Nugent score.

sequences from uncultured bacteria do exist in the NCBI database. In addition to *G. vaginalis*, multiple OTUs were also identified for other species, including *A. vaginae* (two OTUs with 98.8% sequence similarity), *Bifidobacterium breve* (two OTUs with 99.5% sequence similarity) and *Veillonella montpellierensis* (two OTUs with 99.3% sequence similarity). Furthermore, the method was able to differentiate between closely related species or species groups. For example, the sequence for *L. jensenii* could be distinguished from the *L. acidophilus*/*casei*/*crispatus*/*gallinarum* group, despite sharing 97.2% of DNA in the sequenced region. Similarly, *Mobiluncus curtisii* could be differentiated from *Mobiluncus mulieris* (98.1% similarity).

4.4.4 Vaginal microbiome community types in the VMB-HARP population

Hierarchical clustering of the rarefied reads from 871 samples resulted in 37 fine scale clusters, which were pooled into seven biologically meaningful VMB community types for use in downstream clinical epidemiological analyses (see Figure 4.3 for sample clustering tree, Table 4.2 for the definitions of the community types, Table 4.3 for the reasoning behind the choice of community types, Figure 4.4 for the composition of fine scale clusters, and Appendix H for a written description of each fine scale cluster). The following community type descriptions are based on raw read counts. Two types were dominated by lactobacilli, the largest of which was dominated by *L. iners* (type "Li"; n = 214) and the smaller contained predominately *L. acidophilus*/*casei*/*crispatus*/*gallinarum* (which in this niche most likely represents *L. crispatus* and as which it will be referred to from here on) or *L. jensenii* (type "Lcj"; n = 68). Most of the samples of these two types had >50% relative abundance of the dominant lactobacillus species and all had >70% relative abundance of all lactobacilli combined. Most samples of these types also contained very small proportions of BV-associated bacteria, most commonly *G. vaginalis* (median relative abundance 0.08%), and pathobionts (streptococci, staphylococci, *Escherchia*/*Shigella*, *Proteus* and/or *Salmonella*). Pathobionts (i.e. bacteria that have pathologic potential, but may be normally found associated with the healthy human body) were more common in the Li type (99.5% of samples) compared to the Lcj type (65% of samples) but the median relative abundance was very low in both types (0.02% and 0.01%, respectively). Three further VMB community types were characterised by the presence of substantial proportions of bacteria typically associated with BV, including *G. vaginalis*, *A. vaginae*, *BVAB1*, *Megasphaera*, *P. bivia*, *S. amnii*/*sanguinegens* and *V. montpellierensis*. These were further divided

Table 4.2 Vaginal microbiome type definitions used in this study.

VMB COMMUNITY TYPE	DESCRIPTION	CLUSTERS	N (visit 1)	N (visit 5)
Lactobacillus-dominated: *L. crispatus* /*L. jensenii* group ("Lcj")	90% of samples have at least 50% relative abundance of lactobacilli (which may be more than one species) and the proportionally largest OTU is *L. crispatus* or *L. jensenii*.	D1, D2a, G	28	40
Lactobacillus-dominated: *L. iners* ("Li")	90% of samples have at least 50% relative abundance of lactobacilli (which may be more than one species) and the proportionally largest OTU is *L. iners*.	Ca, Cc	117	97
Bifidobacterium-dominated ("BD")	90% of samples have at least 50% relative abundance of *Bifidobacterium* spp	A13, A15	0	2
Lactobacilli with BV anaerobes ("L+A")	90% of samples have at least 10% relative abundance of lactobacilli (which may be more than one species) in combination with bacteria typical of BV	A3, A4a, A4b, Cb, D2b, F2a, F2b	111	97
Bacterial vaginosis type ("BV")	Contains anaerobes typically associated with BV, but none dominate (as per definition above) and there is no significant amount of lactobacillus (as per definition above).	A1a, A1b, A1c, A1e, A2, A6, A7, A8, A9, A10, B1a	152	151
BV anaerobe-dominated ("AD")	90% of samples have at least 50% relative abundance of an anaerobe associated with BV (single species, but may be more than one OTU)	B1b, B2, E, F1, I	29	27
Pathobiont-characterised ("PB")	90% of samples have at least 25% relative abundance of pathobionts (streptococci, staphylococci, Enterobacteriaceae)	A1d, A5, A14, H1, H2, J	8	11

Table 4.3 Reasoning behind the pooling of fine scale clusters into the seven vaginal microbiome types.

VMB COMMUNITY TYPE	REASONING FOR POOLING OF THESE CLUSTERS
Lactobacillus-dominated: *L. crispatus* /*L. jensenii* group ("Lcj")	A *L. crispatus* cluster is commonly encountered in VMB studies using next-generation sequencing (van de Wijgert et al 2014) and the two *L. crispatus*-dominated fine scale clusters closely associated by hierarchical clustering. *L. jensenii*-dominated clusters were rare (consisting of 4 endline samples) and were placed in the same group as *L. crispatus* as these species have been associated with a lower prevalence and increased clearance of HPV (Borgdorff et al 2014, Brotman et al 2014, Dols et al 2012, Reimers et al 2016).
Lactobacillus-dominated: *L. iners* ("Li")	This is a common type, which has been found in the majority of VMB studies using next-generation sequencing (van de Wijgert et al 2014). It consisted of two fine scale clusters which closely associated by hierarchical clustering.
Bifidobacterium-dominated ("BD")	This cluster contained only two samples which were each dominated by a different *Bifidobacterium* sp.
Lactobacilli with BV anaerobes ("L+A")	It has been observed that, compared to other lactobacilli, an *L. iners*-dominated VMB is more likely to transition to a high diversity VMB (Gajer et al 2012, Mitchell et al 2009, Verstraelen et al 2009). Therefore, considering that a VMB consisting of *L. iners* in combination with BV-anaerobes may represent a transitional state, we opted to put these samples in a separate category.
Bacterial vaginosis type ("BV")	This group is the most diverse of the community types in this study and, although it could have been subdivided further, we decided against this because this would have resulted in too small a sample size. The BV type fine scale clusters all closely associated by hierarchical clustering.
BV anaerobe-dominated ("AD")	These samples were mainly dominated by *G. vaginalis* and associated poorly with each other by hierarchical clustering due to the diversity of *G. vaginalis* strains present in them which were differentiated by the OTU clustering tool used in this study. *Gardnerella* dominated clusters have been identified in other studies (van de Wijgert et al 2014). A small number of additional samples (N = 3) were dominated by *A. vaginae*. These were added to this group as both bacteria are commonly associated with bacterial vaginosis.
Pathobiont-characterised ("PB")	Several facultatively anaerobic enteric bacteria have been associated with inflammation and opportunistic infections in the vaginal niche (Donders et al 2017). In this study, these bacteria co-occurred with lactobacilli and BV-associated bacteria and tended to be closely associated by hierarchical clustering with these species rather than each other. However, we have grouped them here due to their association with inflammation and infection. Furthermore, there is some evidence that these clusters may constitute a separate clinical entity coined "aerobic vaginitis" because, in contrast to bacterial vaginosis, they are associated with overt vaginal inflammation on speculum examinations (Donders et al 2017).

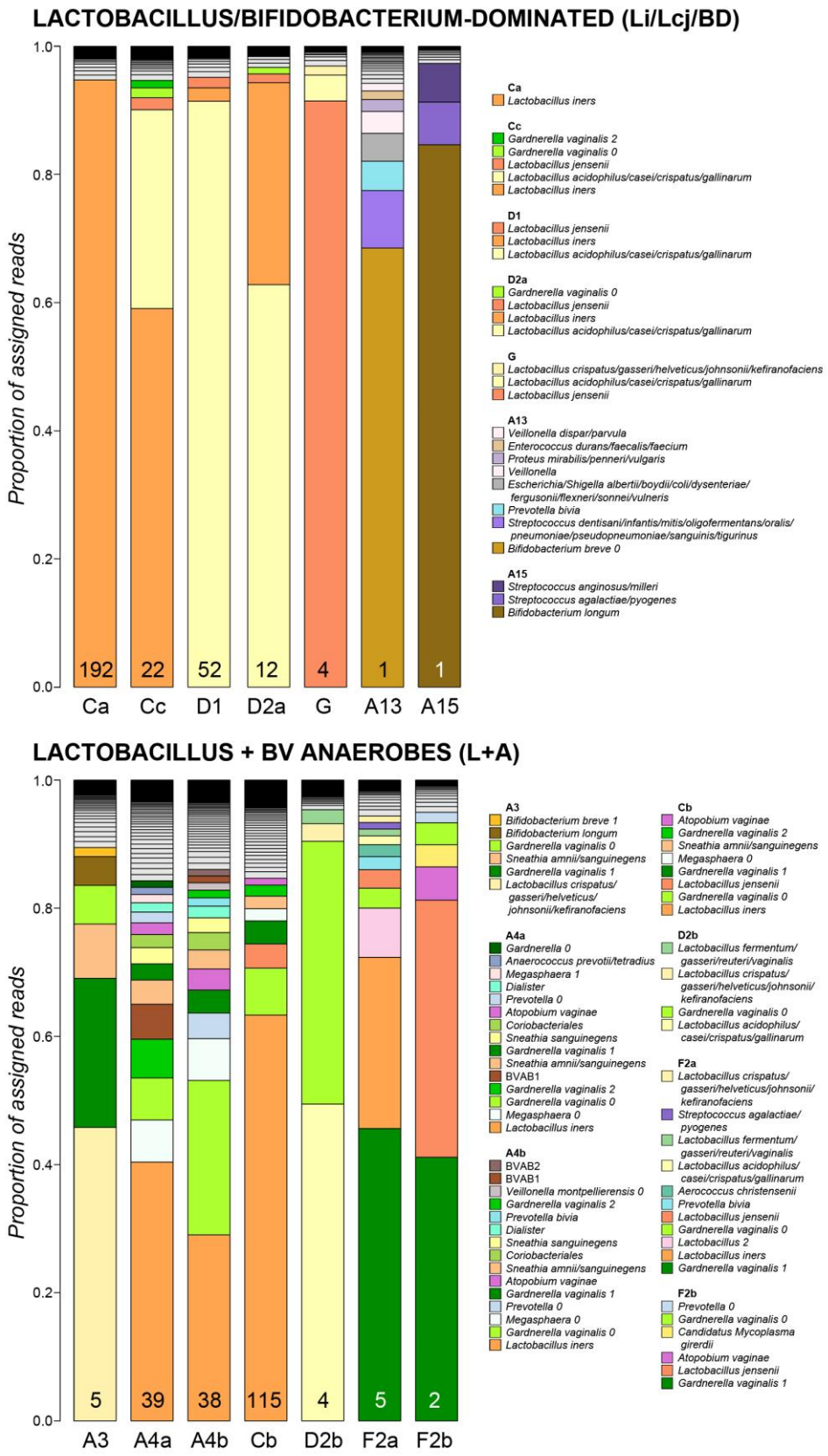


Figure 4.4 Summary barcharts of sample clusters obtained by hierarchical clustering. Samples per cluster are indicated at the bottom of each bar. For clusters containing multiple samples, bars represent the mean of each OTU. Each bar and the corresponding key are arranged in order of abundance. For simplicity, OTUs with a mean relative abundance below 1% are shaded light grey and not represented in the key. Continued on next page.

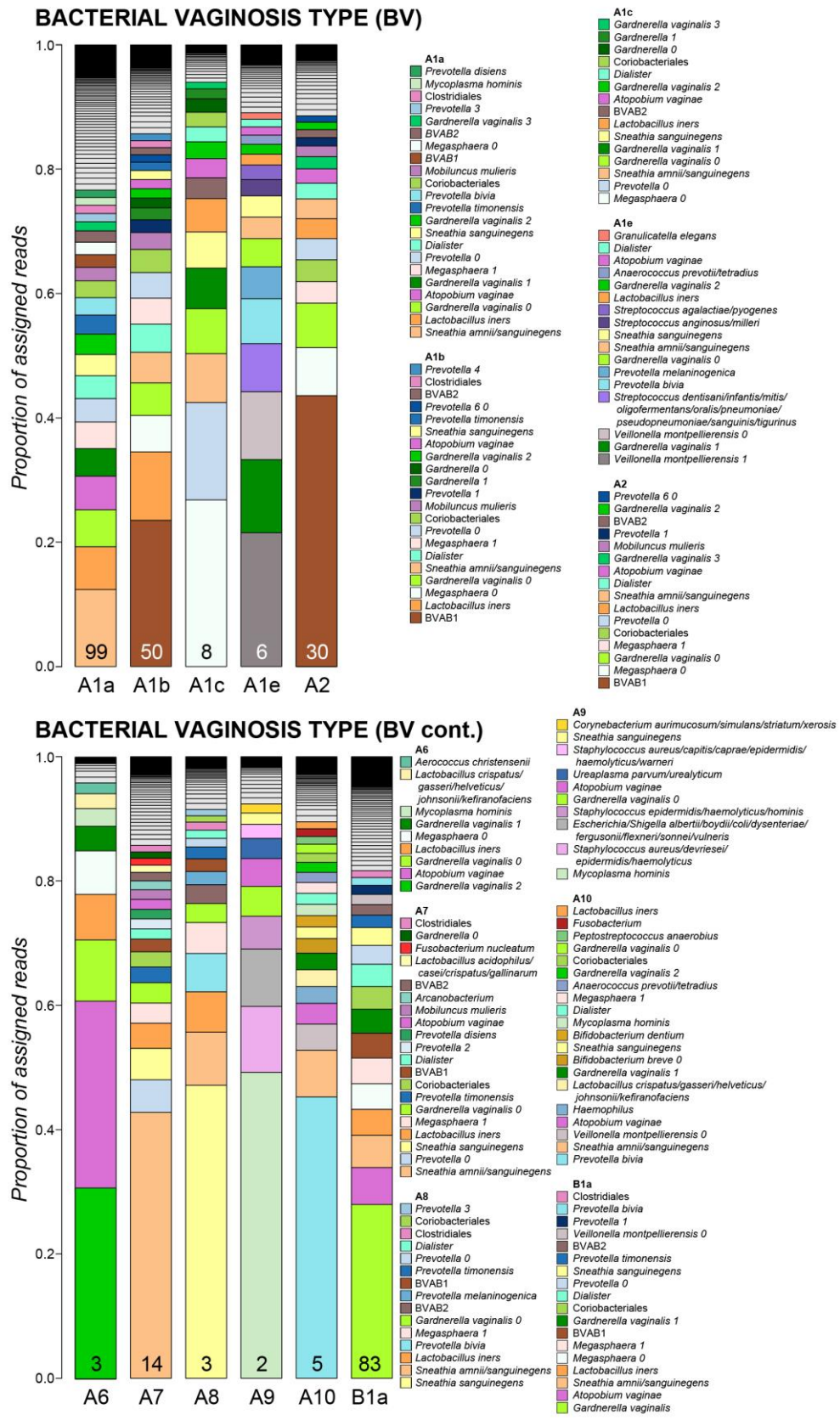


Figure 4.4 Continued on next page.

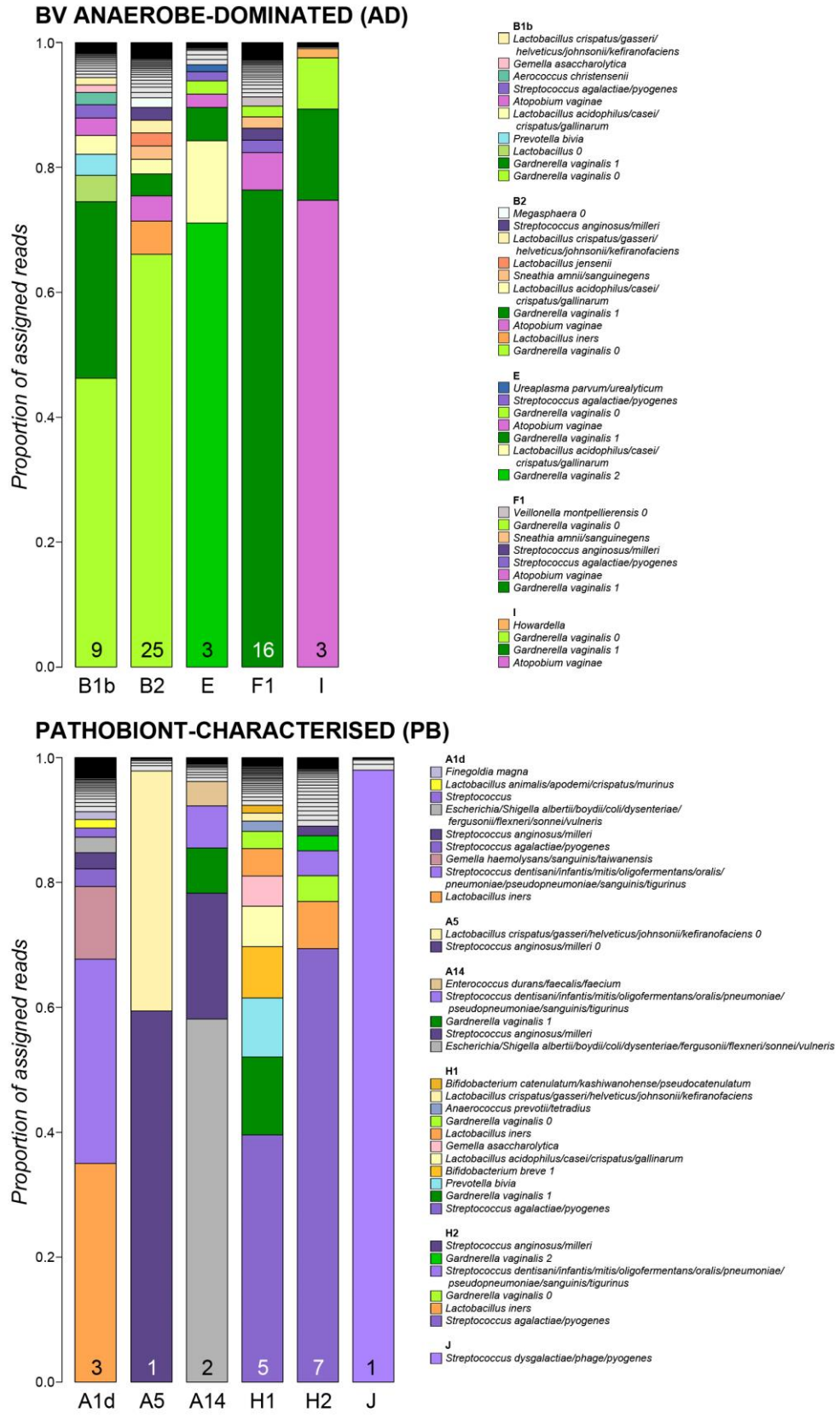


Figure 4.4 Continued from previous page.

into those that were dominated by a single anaerobe species, either *G. vaginalis* or *A. vaginae* (type "AD"; n = 56), those that also contained a sizeable proportion of lactobacilli (90% of samples having at least 10% relative abundance of lactobacilli; type "L+A", n = 208) and those in which no single species dominated (type "BV", n = 303). A further VMB type consisted of samples that contained a substantial proportion of pathobionts (90% of samples have at least 25% relative abundance of pathobionts; type "PB", n = 19) which included *Escherichia*/*Shigella*, *Streptococcus anginosus*/*milleri*, *S. agalactiae*/*pyogenes* and *Streptococcus dysgalactiae*/*pyogenes*. The final type contained only two samples that were dominated by either *Bifidobacterium breve* or *Bifidobacterium longum* (type "BD"). A single sample was dominated by *Scardovia wiggsiae* (a *Bifidobacteriaceae* species reported most commonly from the oral cavity) and was not assigned to a VMB type.

As expected, the lactobacillus percentage was highest for the Lcj and Li community types, intermediate for the L+A type and lowest for the BV, AD and PB types (Figure 4.5). These differences were statistically significant (P <0.0001 in all cases after Holm-Bonferroni correction). Similarly, alpha diversity (Simpson index) was lowest for the Lcj and Li types, intermediate for the L+A, AD, and PB types and highest for the BV type (Figure 4.6). These differences were statistically significant (P <0.001 in all cases after Holm-Bonferroni correction, and additionally the diversity of the L+A type was significantly higher than that of the AD type (P <0.0001).

4.4.5 Correlation of molecular data with bacterial vaginosis by Nugent score

At baseline, women were evaluated for BV by Nugent score and a valid result was available for 435/445 (97.8%) women for which a VMB profile had been obtained at the same visit. As described in section 1.2.1, the Nugent score provides a crude assessment of VMB composition using microscopy and is therefore expected to correlate with sequencing results. The variation in VMB composition was graphed using nonmetric multidimensional scaling (NMDS) in three dimensions, illustrating that the 16S rRNA composition was associated with the Nugent score (Figure 4.7). As expected, the samples with a high Nugent score had a higher alpha diversity (Spearman's rank correlation, P <0.0001) and lower relative abundance of lactobacilli than those with a low Nugent score (P <0.0001; Figure 4.7). Furthermore, the Nugent score results were significantly different between the different fine scale clusters (Figure 4.3) and vaginal community types (Figure 4.8),

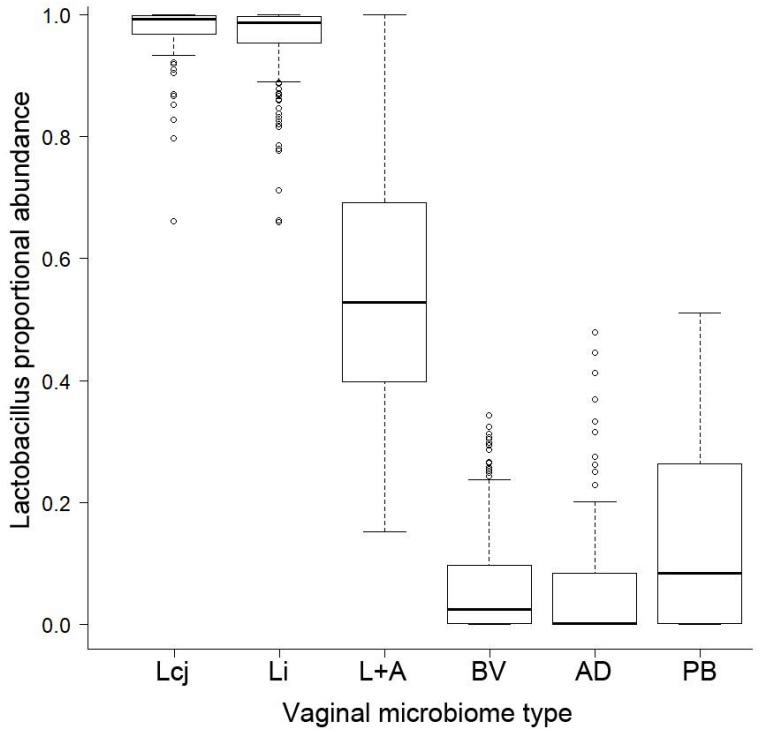


Figure 4.5 *Lactobacillus* relative abundance by vaginal microbiome type. The Bifidobacterium-dominated cluster is not shown due to the low sample size. Lcj: *L. jensenii*/*L.crispatus* dominated; Li: *L. iners*-dominated; L+A: Lactobacilli with BV anaerobes; BV: Bacterial vaginosis type; AD: BV anaerobe-dominated; PB: Pathobiont-characterised.

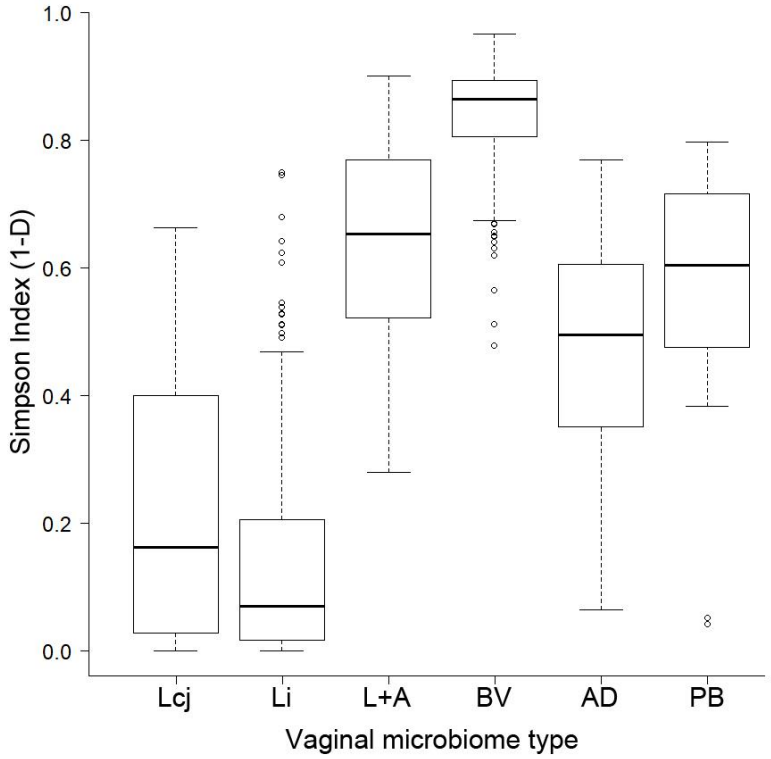


Figure 4.6 Simpson index (1-D) by vaginal microbiome type. The Bifidobacterium-dominated cluster is not shown due to the low sample size. Lcj: *L. jensenii*/*L.crispatus* dominated; Li: *L. iners*-dominated; L+A: Lactobacilli with BV anaerobes; BV: Bacterial vaginosis type; AD: BV anaerobe-dominated; PB: Pathobiont-characterised.

with Li and Lcj having the lowest scores, intermediate for L+A, AD and PB types and highest for BV (Mann Whitney U test, P ≤0.0001 in all cases after Holm-Bonferroni correction). This illustrates the value in 16S rRNA sequencing in being able to distinguish the different species of Lactobacillus and in providing much more detail on the species present in non-Lactobacillus dominated VMB types.

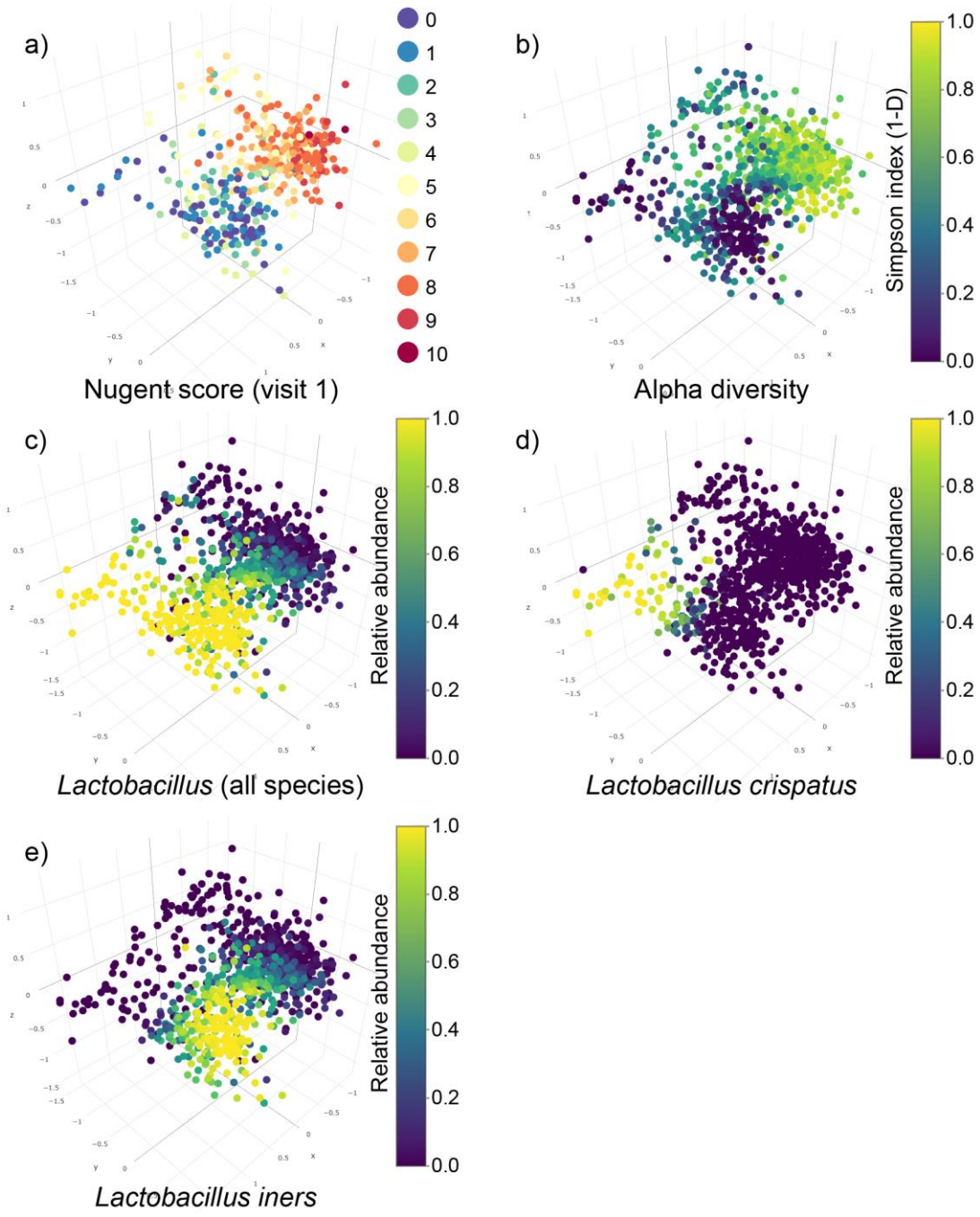


Figure 4.7 NMDS plots summarising the variation in composition between samples in three dimensions. This distribution closely follows the Nugent score (a) and Simpson index (b). The distribution of lactobacilli is shown in graph c, d and e to aid interpretation.

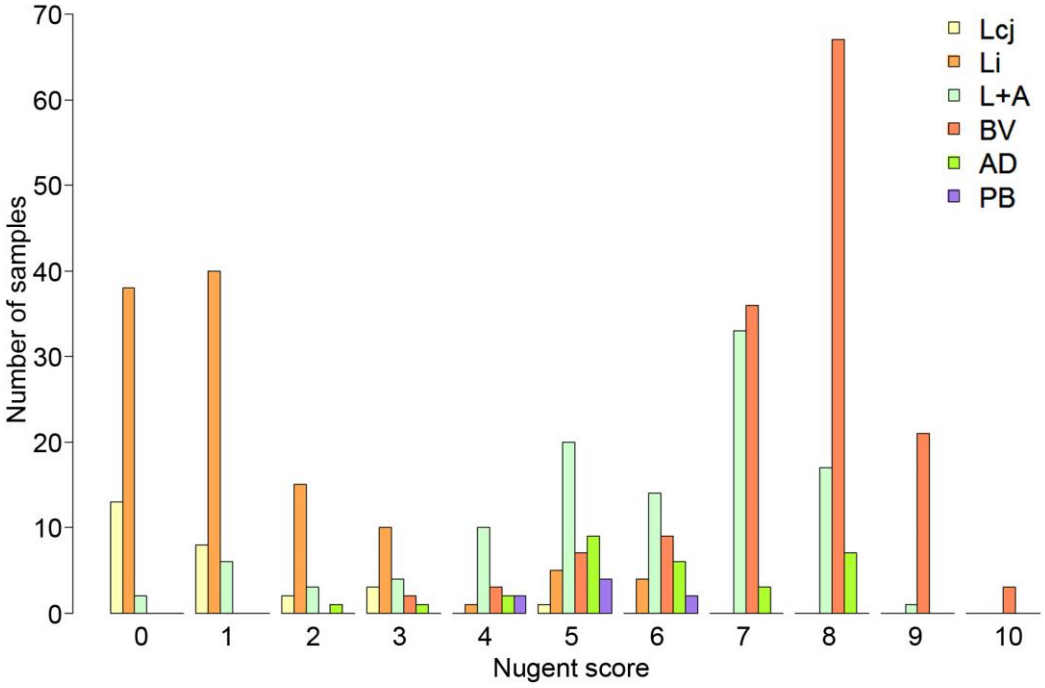


Figure 4.8 Distribution of Nugent scores by vaginal microbiome type at visit 1 (baseline).

4.4.6 Change in VMB composition over time

The VMB of only 143 of the 414 women (35%) who had data available for both visits remained in the same VMB community type between visits 1 and 5. Of the 5 most common VMB types, women who had BV or Lcj at baseline were most likely to have the same VMB type at endline (50.7% and 44.4%, respectively) while those with Li, L+A or AD were less stable (25.5%, 26.2% and 14.8%, respectively). The VMB types most likely to transition to a BV-type microbiome were Li and L+A (31.8% and 31.1%, respectively), while this was least common with Lcj and AD types (11.1% in both cases). The median Bray-Curtis similarity between the baseline sample and that taken at visit 5 was 33% for women with an Lcj type at the first visit, 44% for women with an Li type, and 40% for women with a L+A type. The median Bray-Curtis similarity between visits 1 and 5 samples was lower for women who had a BV (22%), BD (19%), or PB type (6.7%) at visit 1. This difference is statistically significant ($P < 0.001$).

4.4.7 Unadjusted associations between VMB and HR-HPV

In Part 1 of the study – which aimed to investigate the association between the VMB and incidence, persistence and clearance of HR-HPV in HIV-infected South African women – the control group (i.e. HR-HPV negative at both time points) had the highest proportion of the *L. crispatus*-dominated VMB at both baseline and endline (Figure 4.9 and Tables 4.3 and 4.4). When compared to the most common VMB

type (*L. iners*-dominated, Li), this difference was not statistically significant at baseline, but at endline was significantly different for the incident group (odds ratio (OR) = 0.1, P = 0.019) and approached significance for the persistent HR-HPV group (OR = 0.3, P = 0.074). The control group also had the lowest median diversity score at baseline (0.50) and the highest median relative abundance of lactobacilli (73%; Figure 4.9 and Tables 4.3 and 4.4). This difference was statistically significant for the type swap group only (Simpson index OR = 4.5, P = 0.021, *Lactobacillus* relative abundance OR = 0.3, P = 0.008). The control group also had the lowest median diversity score (0.49) at endline and a relatively high *Lactobacillus* relative abundance (64%), but at this time point the *Lactobacillus* relative abundance was slightly lower than in the incident HR-HPV group (69%; Figure 4.9 and Tables 4.3 and 4.4). These differences were not statistically significant, but approached significance for the clearance group (Simpson index OR = 3.3, P = 0.052, *Lactobacillus* relative abundance OR = 0.4, P = 0.097). Both the incident and cleared HR-HPV groups have a lower alpha diversity and a higher *Lactobacillus* relative abundance at the visit at which HR-HPV was present. Non-metric multidimensional scaling in three dimensions did not result in obvious separation between groups (see Figure 4.10a and 4.10b).

LEfSe analysis was performed to compare each of the VMB-HARP study groups to its control group. Discriminant OTUs were found only at endline and are shown in Figure 4.11. We also used LEfSe to compare the HR-HPV status (positive vs negative) for all women who had data available for both visits (including those with CIN2+ at any time point) and found that the median relative abundance of *L. iners* was lower at both visits in women who were HR-HPV positive at that visit.

4.4.8 Unadjusted associations between VMB and CIN2+ in HR-HPV infection

In Part 2 of the study – which aimed to investigate the association between the VMB and the presence of CIN2+ lesions among HIV-infected South African women – the groups had similar proportions of the *L. crispatus*-dominated VMB at both baseline and endline. However, the BV-type VMB was more highly represented in the incident and persistent CIN2+ groups compared to the cleared CIN2+ and control (persistent HR-HPV) groups at both time points (Figure 4.9 and Tables 4.5 and 4.6). This difference was not statistically significant when compared to the most common VMB type (*L. iners*-dominated, Li), but approached significance for the incident CIN2+ group at endline (Table 4.7), which also had a relatively low proportion of

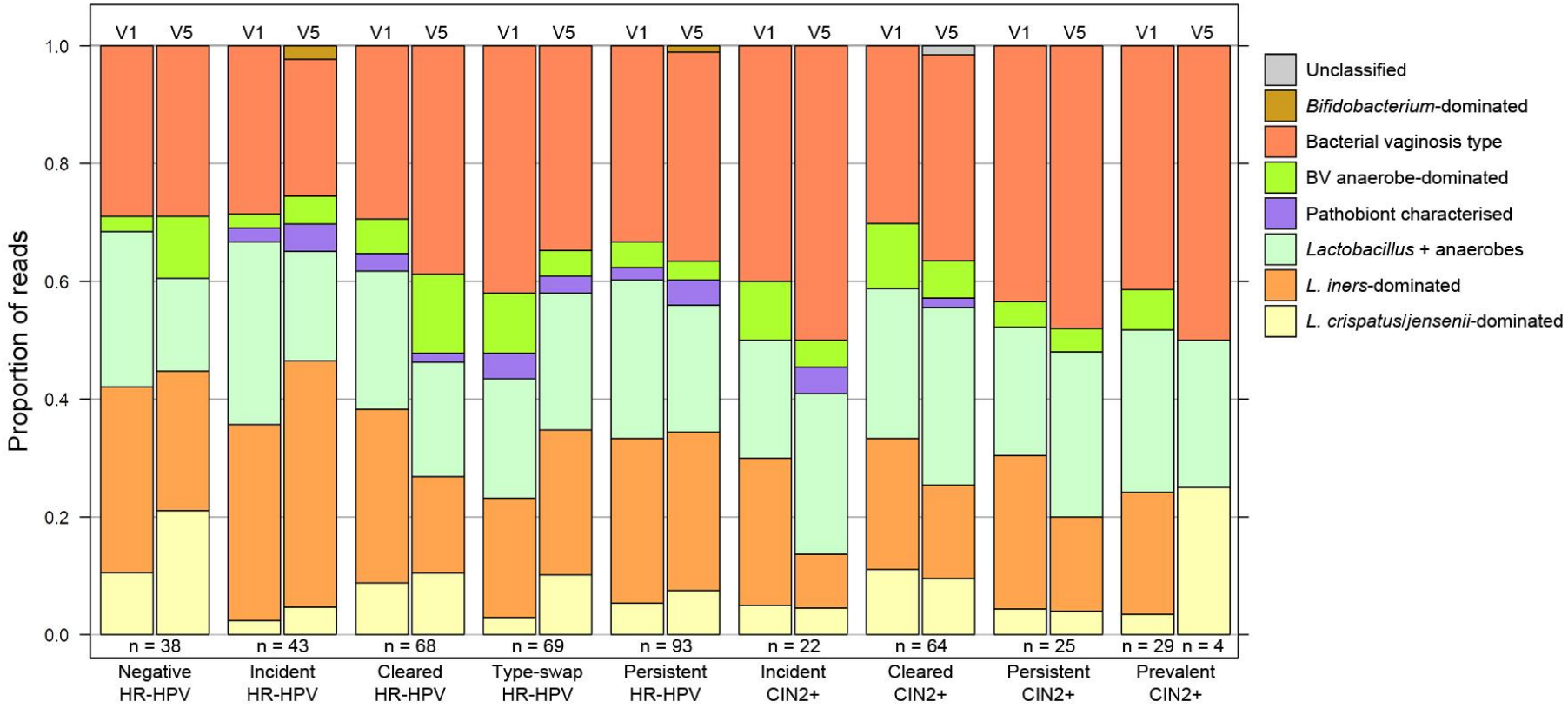


Figure 4.9 Prevalence of VMB types by study group. V1 = visit 1 (baseline); V5 = visit 5 (endline). The number of women in each group is given below each set of bars.

Table 4.4 Results of univariable multinomial logistic regression modelling comparing each VMB-HARP group in Part 1 to the control group at baseline (visit 1). Odds ratios with a P value <0.05 are indicated in bold and starred.

	Negative HR-HPV	**Incident HR-HPV**		**Cleared HR-HPV**		**Type swap HR-HPV**		**Persistent HR-HPV**	
	N (%)	N (%)	OR (95% CI)	N (%)	OR (95% CI)	N (%)	OR (95% CI)	N (%)	OR (95% CI)
Vaginal microbiome type									
Lcj	4 (10.5%)	1 (2.4%)	0.2 (0.0-2.2)	6 (8.8%)	0.9 (0.2-3.9)	2 (2.9%)	0.4 (0.1-2.8)	5 (5.4%)	0.6 (0.1-2.5)
Li	12 (31.6%)	14 (33.3%)	ref	20 (29.4%)	ref	14 (20.3%)	ref	26 (28.0%)	ref
BD[2]	0 (0%)	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)	sparse data
L+A	10 (26.3%)	13 (31.0%)	1.1 (0.4-3.4)	16 (23.5%)	1.0 (0.3-2.8)	14 (20.3%)	1.2 (0.4-3.7)	25 (26.9%)	1.2 (0.4-3.1)
BV	11 (28.9%)	12 (28.6%)	0.9 (0.3-2.9)	20 (29.4%)	1.1 (0.4-3.1)	29 (42.0%)	2.3 (0.8-6.4)	31 (33.3%)	1.3 (0.5-3.4)
AD	1 (2.63%)	1 (2.4%)	0.9 (0.0-15.2)	4 (5.9%)	2.4 (0.2-24.1)	7 (10.1%)	6.0 (0.6-56.0)	4 (4.3%)	1.8 (0.2-18.3)
PB	0 (0%)	1 (2.4%)	sparse data	2 (2.9%)	sparse data	3 (4.3%)	sparse data	2 (2.2%)	sparse data
Total	38	42		68		69		93	
	Median (IQR)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)
Median Simpson Index (IQR)	0.50 (0.15-0.81)	0.67 (0.17-0.83)	1.7 (0.4-6.4)	0.57 (0.24-0.79)	1.5 (0.4-5.0)	0.72 (0.48-0.85)	**4.5 (1.3-16.0)***	0.63 (0.18-0.82)	1.5 (0.5-4.9)
Median lactobacillus relative abundance (IQR)	0.73 (0.20-0.99)	0.48 (0.16-0.95)	0.6 (0.2-1.8)	0.55 (0.05-0.97)	0.6 (0.2-1.5)	0.26 (0.01-0.72)	**0.3 (0.1-0.7)***	0.46 (0.04-0.98)	0.5 (0.2-1.3)

Table 4.5 Results of univariable multinomial logistic regression modelling comparing each VMB-HARP group in Part 1 to the control group at endline (visit 5). Odds ratios with a P value <0.05 are indicated in bold and starred.
†P value <0.1

	Negative HR-HPV	**Incident HR-HPV**		**Cleared HR-HPV**		**Type swap HR-HPV**		**Persistent HR-HPV**	
	N (%)	N (%)	OR (95% CI)	N (%)	OR (95% CI)	N (%)	OR (95% CI)	N (%)	OR (95% CI)
Vaginal microbiome type									
Lcj	8 (21.1%)	2 (4.7%)	**0.1 (0.0-0.7)***	7 (10.4%)	0.7 (0.2-2.7)	7 (10.1%)	0.5 (0.1-1.7)	7 (7.5%)	0.3 (0.1-1.1) †
Li	9 (23.7%)	18 (41.9%)	ref	11 (16.4%)	ref	17 (24.6%)	ref	25 (26.9%)	ref
BD[2]	0 (0%)	1 (2.3%)	sparse data	0 (0%)	sparse data	0 (0%)	sparse data	1 (1.1%)	sparse data
L+A	6 (15.8%)	8 (18.6%)	0.7 (0.2-2.5)	13 (19.4%)	1.8 (0.5-6.6)	16 (23.2%)	1.4 (0.4-4.9)	20 (21.5%)	1.2 (0.4-3.9)
BV	11 (28.9%)	10 (23.2%)	0.5 (0.1-1.5)	26 (38.8%)	1.9 (0.6-6.0)	24 (34.8%)	1.2 (0.4-3.0)	33 (35.5%)	1.1 (0.4-3.0)
AD	4 (10.5%)	2 (4.7%)	0.3 (0.0-1.6)	9 (13.4%)	1.8 (0.4-8.0)	3 (4.3%)	0.4 (0.1-2.2)	3 (3.2%)	0.3 (0.1-1.4)
PB	0 (0%)	2 (4.7%)	sparse data	1 (1.5%)	sparse data	2 (2.9%)	sparse data	4 (4.3%)	sparse data
Total	38	43		67		69		93	
	Median (IQR)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)
Median Simpson Index (IQR)	0.49 (0.10-0.77)	0.53 (0.05-0.74)	1.2 (0.3-4.3)	0.65 (0.39-0.84)	3.3 (1.0-11.2)†	0.59 (0.21-0.80)	1.9 (0.6-6.1)	0.64 (0.17-0.86)	2.2 (0.7-6.9)
Median lactobacillus relative abundance (IQR)	0.64 (0.07-0.99)	0.69 (0.11-0.99)	1.2 (0.4-3.5)	0.22 (0.02-0.95)	0.4 (0.2-1.2)†	0.50 (0.05-0.97)	0.8 (0.3-2.1)	0.37 (0.03-0.98)	0.7 (0.3-1.6)

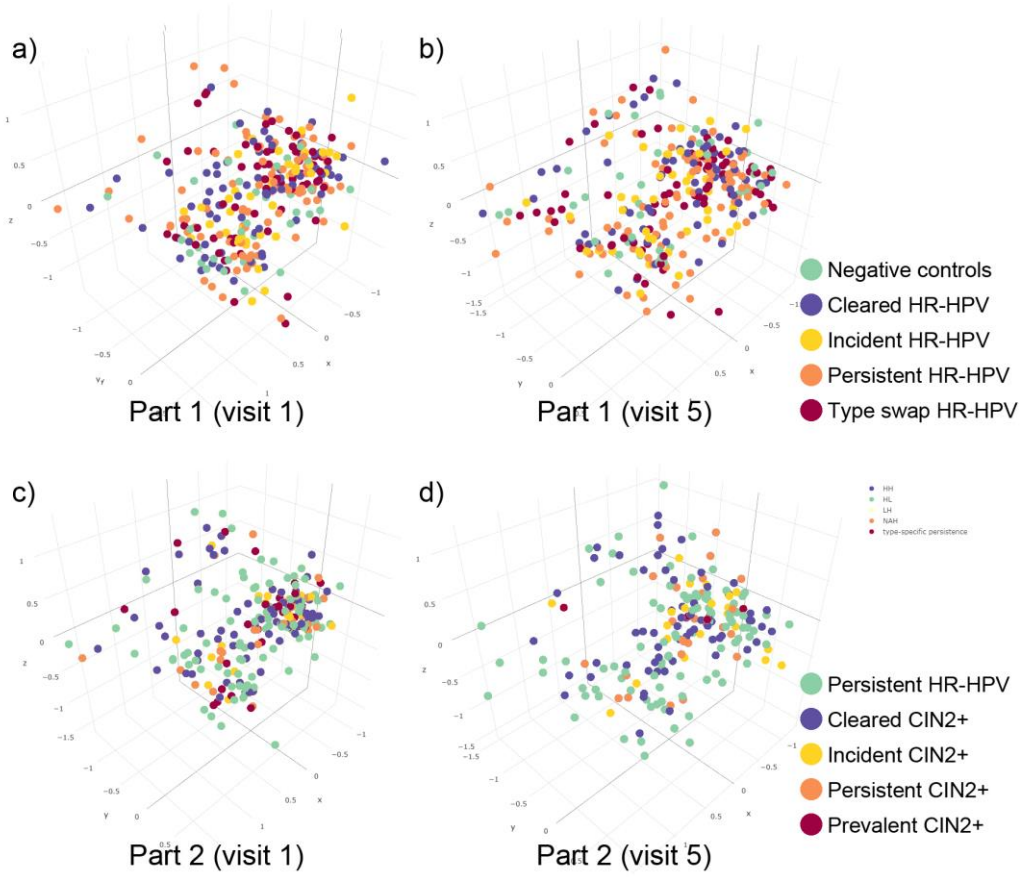


Figure 4.10 NMDS plots summarising the variation in VMB composition between samples in three dimensions. The VMB composition of women with no evidence of CIN2+ (VMB-HARP Part 1) is shown at the top for visits 1 (a) and visit 5 (b). Women with CIN2+ at any time point (VMB-HARP Part 2) are shown in the lower graphs for visit 1 (c) and visit 5 (d), alongside the women with persistent HR-HPV infection but <CIN2 at both time points for comparison.

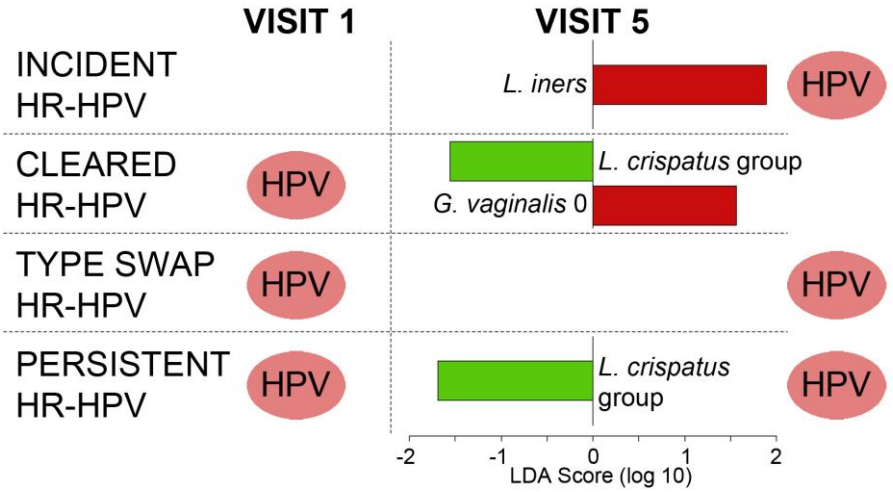


Figure 4.11 Schematic showing OTUs most likely to explain differences between classes listed on the left and negative controls, as identified by the linear discriminant analysis (LDA) score using LEfSe analysis. A positive score shown in red means increased relative abundance and a negative score shown in green a decreased relative abundance. HPV positivity at visits 1 and 5 is also shown.

samples of the Li type (OR 4.2, P= 0.079). At endline, but not at baseline, the L+A type was lowest in the control group (21.5% compared with 27.3%, 30.2% and 28.0% in the incident, cleared and persistent CIN2+ groups, respectively; Figure 4.9 and Tables 4.5 and 4.6). This difference was not statistically significant when compared to the Li type, but approached significance for the cleared CIN2+ group (OR 2.4, P = 0.079).

At baseline, the incident CIN2+ group had the highest diversity score and the lowest *Lactobacillus* relative abundance, with all other groups having similar values, but these differences were not statistically significant. However, at endline, both the incident CIN2+ and the persistent CIN2+ group showed this pattern. The difference in diversity was statistically significant when compared to the control group at endline for the incident CIN2+ group (OR 7.0, P = 0.031) and approached significance for the persistent CIN2+ group (OR 4.0, P = 0.077). Additionally, the difference in *Lactobacillus relative abundance* approached significance for the incident CIN2+ group (OR 0.3, P = 0.093). The results of NMDS in three dimensions are shown in Figure 4.10c and 4.10d. Although no obvious separation of the groups is evident, at visit 5 it does appear that both the incident and prevalent CIN2+ groups are found mainly towards the negative side of the x-axis, which corresponds to the region with higher bacterial diversity and a low *Lactobacillus* relative abundance.

We were unable to identify any OTUs that were significantly different between the VMB-HARP groups by LEfSe. We also used LEfSe to compare CIN2+ status (CIN2+ vs <CIN2) for all women who had data available for both visits, but no differences were identified between these groups either.

4.4.9 Associations of VMB with HR-HPV and CIN2+ in multivariable models

In order to correct for confounding, potential confounding variables were selected *a priori*. These variables were: sexual activity (currently having a regular partner and number of sexual partners in the last 3 months), ART status, log plasma viral load, age, co-infections (chlamydia, candiasis, trichomoniasis and *M. genitalium*) and vaginal cleansing at baseline as well as self-reported smoking, current hormonal contraceptive use of any type and CD4+ T-cell count at both baseline and endline. Pregnancy was not included as pregnant women were excluded from the study at baseline and only one woman was pregnant at the endline visit. Active syphilis and

Table 4.6 Results of univariable multinomial logistic regression modelling comparing each VMB-HARP group in Part 2 to the control group at baseline (visit 1).

	Persistent HR-HPV	**Incident CIN2+**		**Cleared CIN2+**		**Persistent CIN2+**		**Prevalent CIN2+**
	N (%)	N (%)	OR (95% CI)	N (%)	OR (95% CI)	N (%)	OR (95% CI)	N (%)
Vaginal microbiome type								
Lcj	5 (5.4%)	1 (5.0%)	1.0 (0.1-10.9)	7 (11.1%)	2.6 (0.7-9.7)	1 (4.3%)	0.9 (0.1-8.8)	1 (3.4%)
Li	26 (28.0%)	5 (25.0%)	ref	14 (22.2%)	ref	6 (26.1%)	ref	6 (20.7%)
BD[2]	0 (0%)	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)
L+A	25 (26.9%)	4 (20.0%)	0.8 (0.2-3.5)	16 (25.4%)	1.2 (0.5-2.9)	5 (21.7%)	0.9 (0.2-3.2)	8 (27.6%)
BV	31 (33.3%)	8 (40.0%)	1.3 (0.4-4.6)	19 (30.2%)	1.1 (0.5-2.7)	10 (43.5%)	1.4 (0.4-4.4)	12 (41.4%)
AD	4 (4.3%)	2 (10.0%)	2.6 (0.4-18.3)	7 (11.1%)	3.3 (0.8-13.0)	1 (4.3%)	1.1 (0.1-11.5)	2 (6.9%)
PB	2 (2.2%)	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)
unassigned[2]	0 (0%)	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)
Total	93	20		63		23		29
	Median (IQR)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)	Median (IQR)
Median Simpson Index (IQR)	0.63 (0.18-0.82)	0.74 (0.40-0.88)	2.2 (0.5-10.6)	0.57 (0.21-0.77)	0.8 (0.3-2.2)	0.60 (0.26-0.88)	1.4 (0.3-5.9)	0.67 (0.38-0.87)
Median lactobacillus relative abundance (IQR)	0.46 (0.04-0.98)	0.28 (0.02-0.93)	0.6 (0.2-2.0)	0.50 (0.04-0.93)	1.0 (0.5-2.3)	0.41 (0.00-0.97)	0.8 (0.3-2.6)	0.29 (0.06-0.72)

Table 4.7 Results of univariable multinomial logistic regression modelling comparing each VMB-HARP group in Part 2 to the control group at endline (visit 5). Odds ratios with a P value <0.05 are indicated in bold and starred.
†P value <0.1

	Persistent HR-HPV	**Incident CIN2+**		**Cleared CIN2+**		**Persistent CIN2+**		**Prevalent CIN2+**
	N (%)	N (%)	OR (95% CI)	N (%)	OR (95% CI)	N (%)	OR (95% CI)	N (%)
Vaginal microbiome type								
Lcj	7 (7.5%)	1 (4.5%)	1.8 (0.1-22.7)	6 (9.5%)	2.1 (0.6-8.0)	1 (4.0%)	0.9 (0.1-9.3)	1 (25.0%)
Li	25 (26.9%)	2 (9.1%)	ref	10 (15.9%)	ref	4 (16.0%)	ref	0 (0%)
BD[2]	1 (1.1%)	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)	sparse data	0 (0%)
L+A	20 (21.5%)	6 (27.3%)	3.7 (0.7-20.6)	19 (30.2%)	2.4 (0.9-6.2) †	7 (28.0%)	2.2 (0.6-8.5)	1 (25.0%)
BV	33 (35.5%)	11 (50.0%)	4.2 (0.8-20.5) †	22 (34.9%)	1.7 (0.7-4.1)	12 (48.0%)	2.3 (0.7-7.9)	2 (50.0%)
AD	3 (3.2%)	1 (4.5%)	4.2 (0.3-60.9)	4 (6.3%)	3.3 (0.6-17.6)	1 (4.0%)	2.1 (0.2-25.3)	0 (0%)
PB	4 (4.3%)	1 (4.5%)	sparse data	1 (1.6%)	sparse data	0 (0%)	sparse data	0 (0%)
unassigned[2]	0 (0%)	0 (0%)	sparse data	1 (1.6%)	sparse data	0 (0%)	sparse data	0 (0%)
Total	93	22		63		25		4
	Median (IQR)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)	Median (IQR)	OR (95% CI)	Median (IQR)
Median Simpson Index (IQR)	0.64 (0.17-0.86)	0.81 (0.61-0.85)	**7.0 (1.2-41.0)***	0.63 (0.35-0.84)	1.6 (0.6-4.4)	0.77 (0.59-0.84)	4.0 (0.9-19.0)†	0.77 (0.61-0.83)
Median lactobacillus relative abundance (IQR)	0.37 (0.03-0.98)	0.14 (0.03-0.54)	0.3 (0.1-1.2)†	0.37 (0.02-0.81)	0.7 (0.3-1.6)	0.26 (0.01-0.64)	0.5 (0.2-1.6)	0.37 (0.22-0.56)

gonorrhoea were rare (0.7% and 2.2%, respectively) in the VMB-HARP population while a positive HSV-2 serology result was very common (95.4%), so these variables were not included in multivariable models either. The stage of the menstrual cycle and menopause were considered but excluded as these self-reported variables were inconsistent and therefore considered unreliable. Variables included in the final model for Part 1 were: age (continuous), $\log_{10}$ plasma viral load (continuous), number of sexual partners in last 3 months (categorical; none, 1 or more than 1) and *T. vaginalis* PCR results (positive/negative) at baseline and self-reported smoking at either visit (categorical; current smoking declared at baseline or endline: yes/no). Variables included in the final model for Part 2 were: CD4+ T-cell count (continuous), condom use (categorical; over the last 3 months: had no sexual partner, used condoms never/sometimes or always used condoms) and current hormonal contraceptive use (including combined oral contraceptive pills, injectables, or patches) at baseline.

After adjusting for these variables, the associations found in Part 1 of the study remained essentially unchanged. At endline, the control group had a higher proportion of the *L. crispatus*-dominated VMB when compared to the incident group (compared to *L. iners*-dominated VMB type: OR = 0.1, P = 0.016) and persistent HR-HPV group (compared to *L. iners*-dominated VMB type: OR = 0.3, P = 0.084). The control group had lower diversity scores and higher relative abundance of lactobacilli when compared to the type swap group at baseline (Simpson index OR = 3.2, P = 0.085, *Lactobacillus* relative abundance OR = 0.3, P = 0.027), although only the latter remained significant at a 0.05 significance level. The control group also had lower diversity and higher *Lactobacillus* relative abundance compared to the clearance group at endline, with little change in statistical significance (Simpson index OR = 3.2, P = 0.070, *Lactobacillus* relative abundance OR = 0.4, P = 0.056).

In contrast, after adjusting for confounding, the associations found in Part 2 of the study changed considerably. The difference in diversity and *Lactobacillus* relative abundance between the control group and the incident CIN2+ group at endline were no longer significant at the 0.05 level (Simpson index OR 5.2, P = 0.070, *Lactobacillus* relative abundance OR = 0.4, P = 0.196). As in the unadjusted analysis, there were no significant differences in VMB type between outcome groups at the 0.05 significance level, although at endline the higher level of L+A

type and Lcj type in the cleared CIN2+ group approached statistical significance when compared to the controls (OR 3.4 P = 0.083 OR 2.5 P = 0.072).

4.5 Discussion

The fact that BV results in an increased risk of concurrent detection of HR-HPV infection has been well established (Gillet et al 2011, Gillet et al 2012). However, less is known about the temporality, bacterial species and mechanisms behind this association. We found that women who had HR-HPV infection at any time point in the study had a higher diversity VMB with a lower relative abundance of lactobacilli at baseline. After adjustment for confounding variables, this difference was significant for the type swap group such that for each unit increase in the *Lactobacillus* relative abundance, the odds of being in the control group over the type swap group are increased by a factor of 3.3. The reason for this difference being significant only for this group is unknown, although it may be related to a lack of statistical power. It is also possible that there are behavioural factors predisposing this group to multiple HR-HPV infections and concurrent changes in the VMB. However, even after adjustment for the number of sexual partners, a significant difference remained. We also found a similar difference in diversity and *Lactobacillus* relative abundance at endline for the clearance group, although this did not reach statistical significance. In agreement with this observation, at endline, the HR-HPV negative women had a higher proportion of the *L. crispatus*/*L. jensenii*-dominated VMB than all other groups and this difference was significant for the incident HR-HPV group and approached significance for the persistent HR-HPV group. Accordingly, the results of LEfSe analysis found that the incident group was enriched in *L. iners* and the persistent group had a lower abundance of *L. crispatus* when compared to the HR-HPV negative control group. The fact that these two groups show the most significant difference might suggest that the change in the VMB precedes HR-HPV infection. This would agree with the results of two large cohort studies which enrolled both HIV-positive and HIV-negative women and concluded that BV is a risk factor for incident HR-HPV infection, even after adjustment for important confounders (King et al 2011, Watts et al 2005). The observation that BV results in local pro-inflammatory changes (Jespers et al 2017) and disrupts the cervicovaginal barrier (Borgdorff et al 2016b) suggests a possible mechanism for such an association in which loss of epithelial integrity secondary to inflammation provides access to the basal layer of the vaginal epithelium, allowing infection by HR-HPV. However, no definitive conclusions can be drawn on the

temporality of the association from the results of our study, particularly as there was also a trend in the HR-HPV clearance group for a high diversity VMB low in lactobacilli at endline and LEfSe analysis showed a concurrent enrichment in *G. vaginalis* OTU 0 together with a reduced abundance of *L. crispatus* in this group. It is possible that the association may occur in both directions, with disruption in the VMB increasing the risk of HR-HPV and vice versa. Put together, the differences found in this study suggest an association between HR-HPV infection and a high diversity VMB with a paucity of lactobacilli, in particular *L. crispatus*. This result is in agreement with cross-sectional studies that have shown similar results suggesting a link between diverse bacterial communities lacking in lactobacilli and HR-HPV infection (Borgdorff et al 2014, Dols et al 2012, Gao et al 2013, Lee et al 2013, Oh et al 2015). Furthermore, two small longitudinal studies found that a high diversity VMB was associated with a reduced clearance rate of HR-HPV infection (Brotman et al 2014, Di Paola et al 2017) and another longitudinal study showed that *L. crispatus*-dominated VMB reduced the risk of detection of new HPV types (Reimers et al 2016).

In agreement with our results, some of these studies found a stronger reduction in risk of concurrent HR-HPV infection with *L. crispatus* compared to *L. iners* (Borgdorff et al 2014, Brotman et al 2014, Dols et al 2012, Reimers et al 2016). The reasons for this difference remain uncertain. However, a number of potential mechanisms have been suggested. One of these is that an *L. iners*-dominated VMB is more likely to transition to a VMB characterised by a mixture of BV-associated anaerobes when compared to a *L. crispatus*-dominated VMB (Gajer et al 2012, Mitchell et al 2009, Verstraelen et al 2009). This is consistent with our findings in which 31.8% of women with the Li type VMB at baseline transitioned to a BV type at endline, compared to only 11.1% of women who had an Lcj type VMB. This suggests that either *L. iners* is less able to prevent colonisation by BV-associated anaerobes, or otherwise that it is better able to tolerate the environmental conditions associated with BV. Furthermore, there are a number of differences in the metabolic capabilities of *L. crispatus* and *L. iners*. While both species produce lactic acid which results in a low vaginal pH and is generally considered protective, *L. crispatus* is more strongly associated with low pH than *L. iners* (Ravel et al 2011). Furthermore, *L. crispatus* produces predominately the D-isomer of lactic acid which has been associated with increased mucous viscosity and viral trapping (Nunn et al 2015), while *L. iners* produces mainly the L-isomer which may reduce epithelial

integrity by leading to activation of matrix metalloproteinase (Witkin et al 2013). *L. crispatus* has also been reported to be capable of producing bacteriocins that are thought to reduce colonisation success by other species such as *G. vaginalis* (Ojala et al 2014). It should be noted that the sampling interval in this study was relatively long, with a median time interval of 15.9 months. Considering the dynamic nature of the VMB, it is therefore likely that many transitions in between these time points were missed. However, even over this time interval, clear differences in the frequency of transitions were evident, supporting the idea that the VMB may exist in either a relatively stable state or transitions between a small number of alternative equilibrium states (Gajer et al 2012).

Since persistent infection with HR-HPV is a necessary precursor to CIN and the progression to cervical cancer, it is difficult to ascertain whether an observed association between the VMB and CIN is related to the precancerous changes, the persistent viral infection, or both. By utilising longitudinal data, and comparing women with CIN2+ to women with persistent HR-HPV infection that did not develop CIN2+, this study is the first to attempt to separate these effects. However, although we did find some differences in the VMB between women with incident CIN2+ and persistently HR-HPV positive controls, these differences were not robust to adjustment for confounding. After adjustment for confounding, there were no significant differences in VMB measures between the study groups. There was a non-significant trend towards higher levels of Lcj and L+A VMB types at endline in the group that cleared CIN2+, but this difference – assuming it is not a spurious finding – may be merely a consequence of the clearance of HR-HPV. Although previous studies investigating the VMB by molecular methods in relation to CIN have found some differences (Audirac-Chalifour et al 2016, Mitra et al 2015, Oh et al 2015, Piyathilake et al 2016), none of them included any longitudinal data, so these differences could be due to the presence of persistent HR-HPV infection rather than the associated development of CIN. Furthermore, no adjustment for potential confounding variables was made. There are also three published longitudinal studies that have investigated the relationship between BV and the development of cytological or histological precancerous cervical changes. One of these found that BV was a significant risk factor for incident LSIL or HSIL, but also did not adjust for confounding and HIV status was not reported (Engberts et al 2007). A further smaller study in HIV-positive women also found that BV was a significant risk factor for histologically diagnosed CIN, but the association became

non-significant in multivariable analyses (Lehtovirta et al 2008). In contrast, a larger study in HIV-positive women concluded that there was no association between BV and HSIL (Denslow et al 2011). Both studies in HIV-positive women also assessed the risk of persistent BV (defined as BV at baseline and the following visit by Denslow and colleagues and as BV at 50% of visits by Lehtovirta and colleagues) on precancerous cervical changes and found a trend for an increase in risk, which was significant in one of the studies. However, none of these studies adjusted for HR-HPV infection status so any association with BV cannot be distinguished from an association with persistent viral infection.

The overall picture of the VMB obtained in this study is one where the majority of women have either a VMB dominated by *L. iners* (community type "Li"), a VMB consisting of a mixture of anaerobes typically associated with BV (community type "BV") or lie somewhere between these two (community type "L+A"), with lower levels of *Lactobacillus* spp. – most commonly *L. iners* – together with BV-associated anaerobes. The remaining community types were dominated by one of *L. crispatus*, *L. jensenii*, *G. vaginalis*, *A. vaginae* or *Bifidobacterium* spp. or were characterised by the presence of pathobionts and together made up only 14.6% and 18.8% of the study population at baseline and endline, respectively. This is broadly consistent with other studies where the most commonly identified community types consist of one dominated by *L. iners*, one dominated by *L. crispatus* and one made up of a mixture of different anaerobes with low levels of lactobacilli typical of BV (van de Wijgert et al 2014). However, this latter community type contains a very diverse mixture of VMB profiles, with many samples having almost no OTUs in common and this high degree of heterogeneity is suboptimal in terms of allowing researchers to investigate which groups of bacteria are important in affecting health outcomes. As a result, many research groups have further subdivided this group (Brotman et al 2014, Di Paola et al 2017, Gajer et al 2012). The majority of VMB studies have used some form of hierarchical clustering to define community types. Although our study also used this approach, we opted to use a more stringent cut-off that provided a good level of separation of grossly dissimilar samples. However, this resulted in 37 fine scale clusters, of which 23 contained fewer than 10 samples, and this high number would have precluded meaningful downstream analyses. Clusters therefore had to be pooled, which was achieved by making use of knowledge of VMB structure gained from previous studies, as has been previously reported (Borgdorff et al 2017). For example, hierarchical clustering with a less stringent cut-off would

have combined samples containing up to 100% relative abundance of lactobacilli with VMB profiles containing as low as 42% relative abundance of lactobacilli in combination with BV-associated species. Since *L. iners*-dominated VMB is more likely to transition to a high diversity VMB, this suggests that a VMB consisting of *L. iners* in combination with BV-anaerobes may represent a transitional state, we opted to put these samples in a separate category. This allowed us to separate communities that are likely to have different impacts on health. Nevertheless, even at this fine scale level, hierarchical clustering grouped samples that are likely to have very different biological functions, for example there were a small number of samples containing mainly *L. iners* and *L. jensenii* that were clustered with samples containing higher levels of BV-associated anaerobes together with lactobacilli. This has occurred because these sample types were rare and therefore did not fit well with other profiles. However, only a relatively small number of samples were affected and is likely to be no more of a problem than in any other study utilising hierarchical clustering to delineate community types.

As in our study, the *L. iners* dominated VMB type is usually common and in many populations represents the most frequently encountered VMB type (Frank et al 2012, Ravel et al 2011, Zhou et al 2010). In contrast, the prevalence of a *L. crispatus*-dominated VMB is highly variable between populations and may be uncommon (Frank et al 2012, Hummelen et al 2010) or represent the most frequent VMB type present (MacIntyre et al 2015, Mitra et al 2015). In our study, an *L. crispatus*-dominated VMB was only present in 6.3% and 8.5% of women at baseline and endline, respectively and an *L. jensenii*-dominated VMB was rare, present in 0.9% of women at endline only. This relatively low level may be due to the racial background of the HARP cohort, in which most women identified as black. Several studies have shown that, while race does not generally appear to affect the types of VMB profiles encountered, it does influence the frequency of each type. When compared to women with white or Asian background, black populations tend to have a lower proportion of women with *L. crispatus*- and *L. jensenii*-dominated VMB and a higher proportion of mixed high-diversity VMB communities, typical of BV as well as a higher prevalence of BV-associated bacteria in general (Borgdorff et al 2017, Fettweis et al 2014, Ravel et al 2011, Srinivasan et al 2012, Zhou et al 2010). For example, in a sample of healthy non-pregnant North American women of reproductive age 23/104 (22%) of black women had an *L. crispatus*-dominated microbiome, compared to 44/97 (45%) of white women and 24/96 (25%) of Asian

women (Ravel et al 2011). In two other studies conducted in North American populations, the percentage of black women with this VMB type was even lower at below 10% in a group of 960 African-American women (Fettweis et al 2014) and 19/459 (4%) of samples from 16 individuals in another (Gajer et al. 2012). A more recent study in a population of healthy women in Amsterdam found similar results with 17/109 (16%) African-Surinamese women and 15/74 (20%) of Ghanaian women having a *L. crispatus*-dominated microbiome compared to 38/99 (38%) of Dutch women (Borgdorff et al 2017).

A further potential contributing factor influencing the prevalence of different VMB types in our study is the fact that all women were HIV-positive. Several studies have shown that BV prevalence is higher among women with HIV infection (Chehoud et al 2017, Dols et al 2011). While it has also been shown that disturbances in the VMB are a risk factor for HIV acquisition (Gosmann et al 2017, Low et al 2011), it is less clear whether HIV infection itself has an effect on VMB composition. Studies that have used molecular methods to characterise the VMB have found variable results on the association between community types and individual bacterial species and HIV infection, probably because the sample sizes were generally small. However, of those that reported a significant difference or trend between HIV-positive and HIV-negative women, most found that HIV-positivity was associated with lower relative abundance of lactobacilli, in particular *L. crispatus* and/or *L. jensenii*, higher overall bacterial diversity and increased levels of BV-associated bacteria, such as *G. vaginalis*, *A. vaginae* and *Prevotella* (Benning et al 2014, Dareng et al 2016, Demba et al 2005, Dols et al 2012, Gautam et al 2015, Mitchell et al 2013, Pépin et al 2011, Redelinghuys et al 2017, Schellenberg et al 2011). This is consistent with the results of our study in which a high diversity VMB consisting of a mixture of anaerobes was relatively common and a VMB dominated by *L. crispatus* or *L. jensenii* was not. In this study, the prevalence of an *L. crispatus*-domintated VMB type (6.3% and 8.5% of women at baseline and endline, respectively) was relatively low compared to other studies. One study using next-generation sequencing of the V4 region of the 16S rRNA gene to characterise the microbiome of healthy black South African women of reproductive age reported an *L. crispatus*-dominated VMB in 23/236 (10%) of women (Gosmann et al 2017). Other studies conducted in South Africa using culturing techniques or qPCR reported a higher prevalence of *L. crispatus* ranging between 22-25% (Damelin et al 2011, Jespers et al 2015, Pendharkar et al 2013). However, these differences might

be explained by differences in methodology, since in the latter studies presence of *L. crispatus*, rather than dominance, was measured. A further study on South African women showed that women with HIV infection had significantly lower levels of *L. crispatus*, compared to uninfected women (Dols et al 2012).

While the reason for the association between HIV infection and the VMB requires further investigation, it is possible that HIV infection causes a shift in vaginal bacterial community structure. Since this study did not include HIV-negative women, this was not addressed and is one of the limitations of our study. If that were to be the case, and given that HIV infection is also associated with incidence and persistence of HR-HPV infection, it may confound the relationship between the VMB and HR-HPV infection and cervical cancer. By conducting an analysis stratified by HIV status, Dareng and others (Dareng et al 2016) found some evidence that HIV status may actually modify the association between the VMB on HR-HPV infection: as measured by weighted UniFrac, VMB composition differed significantly between HR-HPV positive and HR-HPV negative women, but only among those women that were HIV-negative. HIV status is effectively controlled for in our study by the exclusion of women without HIV. If HIV infection does weaken the association between the VMB and HR-HPV infection, this might explain the relatively small associations found in this study when compared to others with a similar or smaller sample size, highlighting the importance of taking HIV status into account.

4.6 Conclusion

In summary, this population of women at high risk of HR-HPV infection and cervical cancer showed a distribution of VMB types typical of women with black ethnic background and HIV infection. Our results support an association between HR-HPV infection and a high diversity VMB with a paucity of lactobacilli, especially *L. crispatus*. Conversely, we found no evidence of an association of the VMB with CIN, beyond that due to HR-HPV infection. However, the number of CIN2+ endpoints in this study was too small to draw definitive conclusions and this may have been compounded by the fact that all women were HIV-positive. Future studies investigating the relationship between the VMB and HR-HPV should aim to have a larger sample size, take HIV status into account and aim to characterise the VMB using newer, more accurate methods of species definition. Furthermore, studies on the relationship between the VMB and CIN should use appropriate comparison groups to avoid confounding due to persistent HR-HPV infection.

CHAPTER 5: General Discussion

The overall aim of this study was to accurately characterise the vaginal microbiome (VMB) of HIV-positive South African women to determine whether there is an association between the bacteria that inhabit the vaginal niche and high risk human papillomavirus (HR-HPV) infection and the progression to cervical cancer. Although shotgun characterisation of total DNA in vaginal samples would have provided details on the metabolic potential of the bacterial population and may have provided less biased data on proportional abundance, it does require a much greater sequencing effort and is currently prohibitively expensive for large scale epidemiological studies such as VMB-HARP. Therefore the method of choice for this study was to use the 16S rRNA gene as a marker to determine the bacterial taxa present in each sample, which has the additional advantage that public databases contain far more comprehensive information on the 16S rRNA genes of vaginal bacteria, than on their genomes as a whole.

5.1 Optimising 16S rRNA Microbiome Characterisation

Despite its obvious advantages, there are a number of potential biases associated with 16S rRNA sequencing which should not be ignored. Accurate interpretation of results can only be achieved if researchers make every effort to choose protocols that minimise bias by carrying out sufficient validation work and by using appropriate controls. Of course it would not be feasible to explore every potential source of bias in every study and therefore a choice must be made – based on the available literature – as to which parts of the methods are most likely to require validation or optimisation. Unfortunately, even after all effort is made it is inevitable that some bias will remain and it is equally important to be aware of how the choice of methods has affected the results. At the beginning of this study, we therefore set out to answer the following questions relating to the methodology:

1. Are samples stored in BoonFix® at room temperature suitable for characterising the microbiome by 16S rRNA amplicon sequencing?

2. Does the use of bead-beating or enzymes in addition to lysozyme in the DNA extraction process significantly alter the VMB profiles obtained by 16S rRNA amplicon sequencing?

3. Are the microbiome profiles obtained by 16S rRNA amplicon sequencing significantly contaminated by bacteria originating from the extraction kit and if so how can this signal be minimised?

4. What influence does PCR bias have on the microbiome profile obtained?

5. How accurate is operational taxonomic unit (OTU) delineation using sequence read clustering methods based on a similarity threshold and how accurate are taxonomic assignments inferred from the most abundant DNA sequence in these OTUs using RDP classifier? Can newer clustering methods improve on this?

These questions have been addressed in Chapters 2 and 3, and I discuss the implications and limitations of our findings for each of them below.

5.1.1 Storage in BoonFix® at room temperature

The question of whether samples stored in BoonFix® at room temperature are suitable for characterising the VMB by 16S rRNA amplicon sequencing was particularly relevant for the VMB-HARP study, since the samples had already been collected and stored in this way. Even though the Witwatersrand Reproductive Health and HIV Institute (WRHI) research clinic in South Africa where the samples were sourced is well equipped, freezer space is limited and storage in a fixative at room temperature was therefore an attractive option. Vaginal samples stored in BoonFix® have previously been used for VMB characterisation (Dols et al 2011), but there is no published literature comparing microbiome profiles obtained from these samples to profiles obtained from frozen specimens, which are more commonly used. We were able to show that samples stored in BoonFix® produce a VMB profile that is comparable to samples stored frozen at -80°C. We did this firstly by comparing a subset of vaginal swabs stored in BoonFix® with cervical brush samples collected at the same time and stored frozen in PBS-methanol, and secondly by comparing VMB profiles produced from frozen cervico-vaginal lavage samples extracted either directly or first stored for an additional period of 7 months in BoonFix® at room temperature.

One possible limitation of this work is that we were unable to compare VMB profiles from stored samples with those obtained by immediate processing of fresh samples. This is important because studies using faecal samples have reported that freeze-thawing alters community composition when compared to direct processing of fresh samples (Bahl et al 2012, Fouhy et al 2015). However, it is possible that the length of time fresh samples were kept aside prior to processing (up to 4 hours in one study) may have caused the differences found by allowing bacterial division to take place. Furthermore, most studies on faecal sample storage conclude that any effect

on microbiome profiles is small when compared to the differences between different individuals (Bahl et al 2012, Carroll et al 2012, Fouhy et al 2015, Roesch et al 2009, Tedjo et al 2015), making freeze-thawing unlikely to affect the results of large scale microbiome studies in a significant way.

Two recent studies on faecal samples showed that storage at room temperature for 8 weeks in a fixative (95-100% ethanol or OMNIgene Gut kit) produced results that are comparable to processing on the day of collection (Hale et al 2015, Song et al 2016). As far as I am aware, there is currently no published literature on the suitability of samples stored for a longer period at room temperature for 16S rRNA amplicon studies. Our results provide proof of principle that achieving meaningful results using samples stored for several months is possible, which is especially relevant to studies conducted in poor resource settings where freezers are either unavailable or the electricity supply to run such equipment is unreliable.

5.1.2 Bacterial cell lysis efficiency with bead-beating and enzymes

In a study using a mock community, Yuan and others (2012) found that the choice of DNA extraction method biases 16S rRNA microbiome profiles. Their findings suggested that the addition of further enzymatic digestion steps to the Qiagen QIAamp DNA mini kit protocol could improve the accuracy of obtained profiles when compared to the expected result. In particular, protocols including the enzyme mutanolysin (with or without additional enzymes and bead-beating) produced a significantly better approximation of the expected community profile when compared to the other methods. Although these other methods used different commercial extraction kits and the protocols therefore differed in more ways than the addition of the enzymatic digestion step, this does suggest that the pretreatment with lysozyme could be insufficient to satisfactorily lyse all bacteria in a sample. Furthermore, several studies have reported increased extraction efficiency and/or increased proportional abundance of gram-positive bacteria – particularly the hard-to-lyse taxa *Staphylococcus* and *Streptococcus* – in extraction protocols that include a bead-beating step (Abusleme et al 2014, Guo and Zhang 2013, Knudsen et al 2016, Salonen et al 2010, Wagner Mackenzie et al 2015, Willner et al 2012, Yuan et al 2012). However, these extraction protocols again differed in several other ways. We therefore wanted to determine whether vaginal sample profiles extracted with a commercial DNA extraction kit (the Qiagen DNeasy Blood and Tissue kit including the recommended pretreatment with lysozyme), differed significantly if the protocol

was modified by prolonging the lysozyme digestion step, the addition of mutanolysin and lysostaphin to the enzymatic digestion step or the addition of a bead-beating step. However, we found few differences in VMB profiles between protocols and these were greatly outweighed by the biological variation between samples. We therefore concluded that the addition of further steps to the DNA extraction protocol is not necessary when working with vaginal samples. In agreement with our results, most studies comparing different extraction protocols on environmental or clinical samples also reported that differences between extraction methods were smaller than those relating to biological variation (i.e. differences between samples/subjects) (Kennedy et al 2014, Salonen et al 2010, Wagner Mackenzie et al 2015). Although this means that the underlying biological signal is not obscured by the choice of extraction kit, it is advisable to use the same protocol for all samples within a study and take into consideration the differences in extraction protocol when making comparisons between studies.

5.1.3 Contamination in 16S rRNA amplicon studies

It has been demonstrated relatively recently that bacterial contamination of commercial DNA extraction kits results in contaminant sequences in microbiome profiles (Salter et al 2014). It was therefore important to identify any such contaminants in our dataset by the inclusion of negative extraction controls on all sequencing runs. We found several contaminants originating from both Qiagen kits used in this study, the DNeasy Blood and Tissue kit and the QIAsymphony DSP Virus/Pathogen kit. The most common contaminant in both these kits was *Rhodanobacter glycinis*/*terrae*, followed by *Pseudoalteromonas mariniglutinosa*/*prydzensis*/*tetraodonis* and *Vibrio metschnikovii*. These species have been isolated from soil and water samples (Bowman 1998, Jellouli et al 2009, Romanenko et al 2003, Weon et al 2007), making them obvious candidates as reagent contaminants. *Rhodanobacter* has also been reported as part of the human vaginal and semen microbiome, being found at 4.6% in one sample in the vaginal study and at a median abundance of 2.1% in semen (Audirac-Chalifour et al 2016, Weng et al 2014). Both studies used Qiagen extraction kits and neither reported the use of negative controls.

It has been observed that kit contamination is most significant for low biomass samples (Salter et al 2014), and our data supports this finding (see Appendix B). The vaginal niche is a rich source of bacteria and therefore usually yields plenty of

bacterial DNA such that contamination from reagents is trivial. Consistent with this, the sequencing results from the cervicovaginal lavage samples described in sections 2.6 and 2.7 contained only a negligible amount of contaminant sequences. However, vaginal swab samples are likely to yield less DNA which could explain the very low DNA concentration in extracts from BoonFix®-stored samples. Interestingly, although others have found that storage in ethanol-based media may additionally reduce this yield (Hale et al 2015, Vlčková et al 2012), we actually found that storage in BoonFix® for 7 months significantly increased DNA yield. Low DNA yield would have led to poor sequencing results, making it important to optimise this step. We were able to do this by including the vaginal swab in the extraction process during the proteinase K/"buffer AL" digestion step, allowing us to avoid high levels of contaminants in the results. In future it would be prudent for all 16S rRNA amplicon studies to include negative extraction and PCR controls, allowing either the elimination of significant contaminants by laboratory method optimisation or failing that their elimination from the dataset after sequencing.

Apart from contaminants originating from the DNA extraction kit and PCR reagents, we identified another source of contamination in the cervical brush samples stored in PBS-methanol. The most abundant of these was a *Methylobacterium* sp, which was present in 6/8 samples at up to 28.3% relative abundance. This genus is normally associated with soil and water (Dourado et al 2015) and has been identified as a reagent contaminant in this study (see Appendix G) and others (Salter et al 2014). The likely reason for this bacterium to be such a significant contaminant of the PBS-methanol samples is that these aerobic bacteria are able to grow using methanol (Dourado et al 2015) and they probably flourished in the medium before addition of the vaginal samples and subsequent freezing. The fact that these bacteria were contaminants was obvious in our data due to their absence from the paired BoonFix® samples. However, their biology and in particular the ability to metabolise methanol alone raised a red flag. Considering these results it is advisable to treat any bacteria reported as part of the human microbiome but more usually reported in environmental samples as potential contaminants. Researchers should think carefully about the suitability and handling of storage media and use negative storage medium controls to aid in the identification of possible contaminants.

5.1.4 PCR bias in 16S rRNA amplicon studies

Previous work indicated that the PCR step can be a significant source of bias in 16S rRNA studies with perhaps the greatest differences being caused by primer mismatches, where the primer sequence is not an exact match to the template DNA (Brooks et al 2015, Hong et al 2009, Schirmer et al 2015, Tremblay et al 2015). Primer choice may also cause bias due to different amplicon lengths (with longer amplicons being less efficiently amplified) and differences in amplification and annealing efficiency (He et al 2013, Lee et al 2012, Tremblay et al 2015). In order to help understand the bias caused by our primer choice, we compared the profiles obtained from a mock community of six vaginal species pooled either before or after PCR. We were able to show that there was preferential amplification of *Lactobacillus crispatus* and *Prevotella bivia* over *Lactobacillus iners*, *Lactobacillus jensenii, Atopobium vaginae* and *Gardnerella vaginalis*. However, since 16S rRNA amplicon studies determine the relative abundance of bacteria, we cannot draw any conclusions on how primer choice might have altered the relative abundance of taxa in any given vaginal sample because that is dependent on all of the other bacterial species present in that sample. What we have shown is that even where primers are an exact match to the template, PCR bias can result in proportional abundances that differ from those expected and this could not be entirely explained by 16S rRNA copy number, amplicon length or G+C content.

Due to this bias, comparisons between studies using different primer sets should be made with caution, particularly since – in agreement with our results – the PCR step may represent the most significant source of bias in 16S rRNA amplicon studies (Brooks et al 2015, Hong et al 2009, Schirmer et al 2015, Tremblay et al 2015).

5.1.5 Optimising OTU delineation and taxonomic assignments

Classification to the species level is desirable in VMB research, especially for the *Lactobacillus* genus, where many studies have shown there to be differences in terms of the association of different species with health outcomes. For example, several studies have found an association between HIV infection and decreased vaginal lactobacilli in general, but the effect is strongest with *L. crispatus* (van de Wijgert et al 2014). However, identification to species level from 16S rRNA sequencing data alone can be problematic. This is partly due to the fact that most studies have used OTU clustering algorithms that rely on an identity threshold which bin any sequences that share at least that similarity into the same OTU. Therefore,

if the identity threshold is set to 97%, there is potential for sequences that share only 94% sequence identity to be placed in the same OTU (which would happen if both sequences share 97% to the centroid sequence but none of these differences are mutual). The similarity between two sequences from closely related species is often much higher than that (see Appendix E). In Chapter 3, we showed that using an arbitrary similarity threshold for binning sequences into OTUs can result in separate species being merged into a single OTU and in a single species being split into more than one OTU. This occurred for example with reads from *L. jensenii*, some of which were binned with *L. crispatus* reads, and with reads from *Lactobacillus amylovorus* which were all binned together with *L. crispatus* reads. Although we have shown that this inadequate species separation does not generally alter the representative sequence obtained for the major species in the dataset, it does mean that less abundant species are more likely to be subsumed into OTUs consisting of other species, making the species IDs of any one sequence within a particular OTU far less certain than the ID obtained for that OTU's reference sequence. We also showed that using newer OTU clustering methods such as DADA2 and Swarm considerably improves OTU delineation, thereby greatly improving the accuracy of species identification from 16S rRNA amplicon sequencing data. Although we confirmed that DADA2 can separate DNA sequences that differ by only a single nucleotide, we found that in some cases read error was insufficiently dealt with such that the *L. jensenii* and *L. iners* controls that should have consisted of a single OTU also contained multiple smaller erroneous OTUs. We therefore recommend the use of Swarm, which can accurately separate DNA sequences that differ by at least 2 nucleotides.

In addition to OTU binning, the accuracy of taxonomic assignments of reads is critically dependent on the bioinformatics programme used to assign taxonomy. One such programme that is commonly used in VMB studies is RDP classifier. The potential for error with this tool was nicely illustrated by the results of Chapter 3 where *Bacillus subtilis* in the Zymo Microbial standard was incorrectly identified as *Bacillus mojavensis* in two of the pipelines. When the representative DNA sequence for this bacterium is input into the *assignSpecies* function in DADA2 and searched against the Silva v. 128 database with multiple matches enabled, an exact match is found to a total of nine different species, including *B. mojavensis* as well as the correct species, *B. subtilis*. In both cases where the sequence was incorrectly identified by RDP classifier, the bootstrapping confidence value was comparatively

low (0.80 and 0.86). At this level of accuracy, RDP classifier may misidentify up to 30.8% of species (Wang et al 2007). Despite this, the confidence level of assignments is often not reported in microbiome studies that have used RDP classifier, nor is the cut-off used to make a positive identification, with the default often being as low as 0.50, at which level the accuracy of taxonomical assignments can fall below 50%. For researchers who are not familiar with the species identification methods used, this leads to misplaced confidence in taxonomic assignments. The problem of species identification can be illustrated further by considering the case of the fourth most common *Lactobacillus* OTU in the VMB-HARP dataset, which was an exact match to five *Lactobacillus* species in the Silva v. 128 database according to the *assignSpecies* function: *L. crispatus*, *Lactobacillus gasseri*, *Lactobacillus helveticus*, *Lactobacillus johnsonii* and *Lactobacillus kefiranofaciens*. RDP classifier identified this OTU as *L. gasseri* with a high bootstrap value of 0.99. These results could be explained by two scenarios: the first is that there are indeed five species of *Lactobacillus* encompassing one or more strains that all share an identical DNA sequence in the V3-V4 region and RDP classifier has identified only one of them. However, it is also possible that there are inaccuracies in the Silva database, whereby some DNA sequences are incorrectly labelled. This could lead to some species being listed as possible IDs by *assignSpecies* when they should not be, resulting in unnecessarily ambiguous species assignments. However, in consideration of the fact that DADA2's *assignSpecies* function provided a more satisfactory result in the case of *B. subtilis* and that RDP classifier was inconsistent in its ability to assign the *L. jensenii* control to species level in different pipelines, (despite the reference DNA sequence being the same), we recommend the use of DADA2's *assignSpecies* function in addition to or instead of the species IDs provided by RDP classifier.

Ultimately it is up to individual research groups to decide how to report taxonomic assignments by finding a balance between accuracy, clarity and comparability with other studies. Accuracy of taxonomic assignments with DADA2 depends critically on the accuracy and completeness of the 16S rRNA sequence database when searching for exact matches to the reference sequence with *assignSpecies*. The choice of database also affects the results with RDP classifier, which may be particularly important when working on less well studied microbial niches that may be poorly represented in public databases (Newton and Roeselers 2012). By comparison, the differences in the taxonomies associated with each database

appear to have less of an effect (Werner et al 2012). For the VMB-HARP study, we chose to use the current version of the Silva database, which has the advantage that it is regularly updated and that all sequences are quality checked. Although Silva contains fewer sequences than the Greengenes database, this is likely to be less of a problem for well-studied environments such as the vaginal niche.

Finally it must be noted that even with improved bioinformatics methods, it may not be possible to call species definitively, as it would require each species to have a unique DNA sequence in the region of the 16S rRNA gene being sequenced and this is often not the case (see Appendix E). For example, in the V3-V4 region, the DNA sequence for *L. crispatus* is identical to *Lactobacillus gallinarum*. Although the latter is not associated with the vaginal niche, it is evident that *L. crispatus* cannot be conclusively identified based on its DNA sequence in the V3-V4 region alone. Despite this, VMB studies using this region often refer to an *L. crispatus* OTU, without any mention that this ID entails a degree of uncertainty.

5.1.6 The use of 16S rRNA sequencing to study the vaginal microbiome

This thesis has demonstrated that methods optimisation is important to achieve accurate results in 16S rRNA amplicon studies and extract as much information as possible from the resulting data. Some remaining bias cannot be avoided, of which the choice of primers has perhaps the greatest impact. It is therefore important that researchers are aware of this potential bias and clearly state which primers have been used in their work by not burying this information in the methods section. While greater uniformity in methods within the VMB community would be desirable in order to maximise the comparability of studies, this is currently difficult to achieve as we are in an era during which genomics and bioinformatics methods are constantly evolving and improving. Alternatively, research groups should strive to optimise their own techniques and make use of improved methods when they become available in order to produce accurate, future-proof data. Only then can we hope to answer the outstanding questions in VMB research and use this information to improve human health.

5.2 The Vaginal Microbiome, HR-HPV and Cervical Cancer

5.2.1 The vaginal microbiome and its association with HR-HPV infection

The majority of the research on the association between the VMB and HR-HPV infection has been carried out using data on bacterial vaginosis (BV), diagnosed by

Nugent or Amsel score. From these studies, there is good evidence that there is a positive association between vaginal dysbiosis and HR-HPV infection (Gillet et al 2011, Gillet et al 2012). Although most studies were cross-sectional such that conclusions cannot be drawn on the temporality of this association, there is some evidence from two large multicentre cohort studies that BV (defined as a Nugent score of 7 or higher) is a risk factor for incident HR-HPV infection (King et al 2011, Watts et al 2005). One of these studies also found that BV slightly increased the risk for delayed clearance of HR-HPV infection (King et al 2011). Importantly, there was extensive control in both studies for HIV-related and other covariates and the associations between BV and HR-HPV remained after adjustment for these. However, both studies enrolled large numbers of HIV-positive and a smaller number of HIV-negative women at high risk of HIV infection, which could affect the generalisability of the results. Little is known about whether there is also an increased risk of incident BV after HR-HPV infection.

We have shown that the data from 16S rRNA sequencing correlate well with the Nugent score used for the diagnosis of BV and this concurs with the findings of others (Ravel et al 2011). However, it is also evident that the Nugent score does not provide the level of detail that can be revealed by 16S rRNA sequencing and which is likely to be relevant for health outcomes. For example, it is increasingly recognised that different lactobacilli have different associations with health (van de Wijgert et al 2014) and these species are not distinguished by the Nugent score. By using 16S rRNA sequencing we were able to describe the VMB in the VMB-HARP cohort in much more detail.

In doing so, we have found an association between HR-HPV infection and a VMB that is made up of a diverse population of bacteria and is low in lactobacilli. This is broadly in agreement with the findings of other studies (Borgdorff et al 2014, Brotman et al 2014, Di Paola et al 2017, Dols et al 2012, Gao et al 2013, Lee et al 2013, Oh et al 2015, Reimers et al 2016). However, considering that previous studies have found convincing associations between the VMB and HR-HPV infection despite having a much smaller sample size (Brotman et al 2014, Di Paola et al 2017, Lee et al 2013), the associations found in our study were smaller than expected. One possible explanation for this is that the VMB-HARP study was restricted to HIV-positive women which is likely to have reduced the size of the HR-HPV negative control group, since HIV positive women have a higher prevalence

and persistence of HR-HPV infection (King et al 2011, Watts et al 2005). Concurrently, the number of women with a *L. crispatus*-dominated microbiome was small. Furthermore, there is some limited evidence to suggest that HIV infection could modify the association between the VMB and HR-HPV infection (Dareng et al 2016). To date there are only four molecular studies on the VMB and HR-HPV infection that have taken HIV status into account (Borgdorff et al 2014, Dareng et al 2016, Dols et al 2012, Reimers et al 2016), but the sample sizes of these studies were too small to definitely determine whether HIV infection was an effect modifier or confounder. However, the aforementioned longitudinal cohort studies that investigated the relationship between BV and HR-HPV both concluded that while HIV infection was associated with changes in the VMB, it did not modify the determined odds ratios between BV and HR-HPV (King et al 2011, Watts et al 2005). Future studies should aim to include a large enough sample size to include HIV as a potential confounder or effect modifier, or otherwise limit themselves to HIV-positive or HIV-negative women, a choice that will depend on the study objectives.

The results of the VMB-HARP study also indicate that an *L. crispatus*-dominated microbiome carries a lower risk of concurrent detection of HR-HPV than does an *L. iners*-dominated microbiome, which is in agreement with previous studies (Borgdorff et al 2014, Brotman et al 2014, Dols et al 2012, Reimers et al 2016). We also showed that in comparison to an *L. crispatus*-dominated microbiome, an *L. iners*-dominated microbiome or one containing a smaller proportion or *L. iners* together with BV-associated anaerobes more often transitions to a microbiome typical of BV. This agrees with the findings of others (Gajer et al 2012, Mitchell et al 2009, Verstraelen et al 2009) and may at least in part explain why *L. crispatus* has been found to have a stronger association with positive health outcomes than *L. iners*. However, the mechanisms behind these differences are still poorly understood and warrant further investigation.

5.2.2 Is the vaginal microbiome associated with precancerous cervical changes?

Interesting results from the field of gastrointestinal research suggest that the microbiota can play an important role in the development of cancer. Recent work has suggested that metabolites produced by members of the gastrointestinal microbiome can either promote cancer development or suppress inflammation and

carcinogenesis (Louis et al 2014). Additionally, specific bacterial species have been implicated in certain cancers. In particular, *Helicobacter pylori*, a common member of the gastrointestinal microbiome, has been linked to the development of gastric ulcers and cancer. It is currently the only bacterium recognised as carcinogenic to humans by the IARC (2018) and is thought to cause cancer by manipulating signalling pathways that promote inflammation and cell proliferation and survival (Gagnaire et al 2017).

Similar mechanisms could be at play in the development of cervical cancer, and might be one explanation as to why some women with HR-HPV develop cervical cancer while most do not. In line with this, several studies have suggested a relationship between the VMB and cervical intraepithelial neoplasia (CIN), the precursor to cervical cancer (Audirac-Chalifour et al 2016, Engberts et al 2007, Lehtovirta et al 2008, Mitra et al 2015, Oh et al 2015, Piyathilake et al 2016). Additional limited evidence for a role of lactobacilli in reducing the risk of cancer is provided by *in vitro* work which has found that certain *Lactobacillus* spp. can be cytotoxic to gastrointestinal cancer cell lines (Choi et al 2006, Russo et al 2007). A more recent report found that *L. gasseri* and *L. crispatus* homogenates inhibited the growth of cancerous cervical cells when compared to normal cervical cells, but also reported an anti-apoptotic effect on the cancer cells (Motevaseli et al 2013). A further study found that live *Lactobacillus delbrueckii* did not have an apoptotic effect on cervical cancer cells *in vitro*, but did suppress cell migration which might suggest it could inhibit the formation of invasive cervical cancer. The authors also found a concurrent upregulation of E-cadherin expression in the presence of lactobacilli, which has been associated with a better prognosis in cervical cancer (Li et al 2017). In contrast, Kim and colleagues (2015) found no effect of *Lactobacillus casei* extracts on the growth of cervical cancer cell lines. While these studies suggest that there is potential for lactobacilli to influence the development and progression of cervical cancer, the biological relevance of their results is still unclear.

Research on the role of the VMB in cervical cancer is still in its infancy and very few conclusions can be made from the data that is currently available. In particular, most epidemiological studies are cross-sectional and the small number of published longitudinal studies made no adjustment for HR-HPV infection status, which is likely to have confounded the results. Our study is the first longitudinal molecular study to

investigate the association between the VMB and CIN and to attempt to use women with persistent HR-HPV infection that did not develop precancerous lesions as controls. However, we found no convincing association between the VMB and precancerous cervical changes beyond that which may be due to HR-HPV infection. Larger well-designed longitudinal studies are needed to investigate this question further. Due to the slow progression of initial HR-HPV infection to cervical cancer, women should ideally be followed for a period of several years. Furthermore, in order to distinguish between the effects of HR-HPV and CIN, the control group should consist of women that are also HR-HPV positive. Finally, the consideration of potential confounders is vital, as there are several socio-demographic and health-related variables that may potentially affect both VMB composition and HR-HPV infection. For example, HIV infection is associated with an altered VMB that has higher bacterial diversity, increased levels of BV-associated bacteria and is lacking in lactobacilli, in particular *L. crispatus* and *L. jensenii* and this is consistent with our findings in the VMB-HARP study (Benning et al 2014, Dareng et al 2016, Demba et al 2005, Dols et al 2012, Gautam et al 2015, Mitchell et al 2013, Pépin et al 2011, Redelinghuys et al 2017, Schellenberg et al 2011). Concurrently, women infected with HIV have a higher prevalence of HR-HPV which is probably a consequence of HIV-induced immunosuppression allowing the human papillomavirus (HPV) to persist and this in turn is a risk factor for the development of cervical cancer (Denny et al 2012, Moscicki et al 2004). It is therefore important that HIV status is not ignored in any future studies on the VMB and its association with cervical cancer since it could be an important confounder.

5.2.3 Study limitations and future directions

One of the limitations of this study, the small sample size of the HR-HPV negative control group, has already been mentioned. Additionally, our study was limited to HIV-positive women, which was hoped to increase statistical power, but also means that the results may not apply to the general population. Furthermore, although the study followed women over a median of 15.9 months, samples were only available from baseline and endline. More frequent sampling would have been ideal since HPV detection can be variable (Liu et al 2014) and the VMB is dynamic and may change over relatively short periods of time (Gajer et al 2012). The average time to clear an HPV infection reported in the literature is variable, ranging from 4-20 months (Denny 2009). With a relatively long sampling interval it is therefore possible that some HPV infections classed as persistent may have been reinfections and a

shorter sampling interval of 4-6 months would have been ideal. Future work on the association of the VMB with HR-HPV and CIN should therefore encompass large cohort studies following women over a sufficiently long period of time and carrying out sampling at multiple time points. Furthermore, the control of potential confounders, including HIV and other STIs and socio-demographic variables is required to definitively determine if any association found is real.

A further limitation of our study and other marker gene studies is that they provide limited information on how the VMB community functions, since they do not provide information on the rest of the genome, or what the bacteria are doing. Recently, a comparative genomics study of *L. crispatus* found that the bacterium encodes a protein with the potential to reduce the ability of *G. vaginalis* to adhere to epithelial cells *in vitro* (Ojala et al 2014). A further approach to investigate community function is to study gene expression using transcriptomics and proteomics techniques. Only a few studies have used these techniques so far but they can provide interesting insights into how the bacteria in the VMB interact. For example, using meta-RNA-seq, Macklaim and others (2013) showed that the expression profile of *L. iners* markedly changes in the presence of BV and Borgdorff and others (2016a) used proteomics to demonstrate that *L. iners* glycolytic enzymes are significantly enriched under BV conditions. These studies highlight the potential for functional genomics studies in investigating the metabolic states of bacteria in the vaginal milieu.

However, future studies need also to consider the influence of the host on the vaginal environment and its contribution to the metabolites present. For example, oestrogen is thought to have a strong influence on the VMB, with major shifts in bacterial populations occurring at puberty and menopause (Linhares et al 2011) and the use of hormonal contraceptives is associated with a reduced risk of BV (Van de Wijgert et al 2013). An obvious further means for cross-talk between the human host and the VMB is the mucosal immune system of the lower female genital tract, which is also under the influence of oestrogen (Wira et al 2000). The high diversity microbiome typical of BV has been associated with increased levels of pro-inflammatory cytokines (Gautam et al 2015, Jespers et al 2017), indicating that there is a reaction to the presence of BV-associated microbes from the host. Gaining a better understanding of the vaginal microenvironment, and in particular elucidating the cross-talk between the human host and the VMB, as well as which

factors disrupt the balance between the two, should therefore be a priority for VMB research in the future.

Nonetheless, the current study provides important information on the VMB associated with HPV infection and carcinogenesis, thereby providing a first step towards the development of novel and cost-effective screening and control strategies for the prevention of cervical cancer. Understanding which bacterial communities and/or species are involved in HPV infection and carcinogenesis will allow a targeted approach, for example by more frequent screening of high-risk individuals and by aiding the development of personalised prevention strategies. One such prevention strategy might be the development of effective probiotic treatments. Recent studies have shown that there is potential for probiotics to alter the vaginal flora and reduce the risk of STIs, including HR-HPV infection (Palma et al 2018). However, there is currently a paucity of good quality randomised controlled trials on the use of probiotics to prevent urogenital infections and other reproductive disorders and recent systematic reviews have concluded that there is only weak evidence for their use in the prevention of urinary tract infections (Schwenger et al 2015), vulvovaginal candidiasis (Xie et al 2017) and preterm labour (Jarde et al 2018) such that they cannot currently be recommended for clinical use in place of drug treatments. Clearly, more research is needed before recommendations on the use of probiotics to optimise health through manipulation of the VMB can be made. One of the challenges in making advances in this area is that different studies differ in the bacterial strains or species contained in the probiotics used and many have been carried out using species that are not normally associated with the vagina in any great number. This may significantly affect the results of clinical trials, particularly considering that several epidemiological studies have found differences in health benefits between different species of naturally occurring vaginal lactobacilli (van de Wijgert et al 2014). Determining which bacteria are associated with a reduced risk of HPV infection and carcinogenesis will aid in the choice of probiotic strains. In conclusion, research on the VMB has great potential for improving women's health. However, much more research is needed before these strategies can be implemented on a large scale with sufficient confidence in their accuracy and/or efficacy.

References

Aagaard K, Riehle K, Ma J, Segata N, Mistretta T-A, Coarfa C *et al* (2012). A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *Plos One* **7:** e36466.

Abusleme L, Hong B-Y, Dupuy AK, Strausbaugh LD, Diaz PI (2014). Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing. *Journal of oral microbiology* **6:** 23990.

Acland A, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C *et al* (2014). Database resources of the national center for biotechnology information. *Nucleic acids research* **42:** D7.

Adamowicz MS, Stasulli DM, Sobestanovich EM, Bille TW (2014). Evaluation of methods to improve the extraction and recovery of DNA from cotton swabs for forensic analysis. *Plos One* **9:** e116351.

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C *et al* (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology* **12:** R18.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *Journal of molecular biology* **215:** 403-410.

Amsel R, Totten PA, Spiegel CA, Chen KC, Eschenbach D, Holmes KK (1983). Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. *The American journal of medicine* **74:** 14-22.

Anahtar MN, Byrne EH, Doherty KE, Bowman BA, Yamamoto HS, Soumillon M *et al* (2015). Cervicovaginal bacteria are a major modulator of host inflammatory responses in the female genital tract. *Immunity* **42:** 965-976.

Anderson MJ (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology* **26:** 32-46.

Anderson MR, Klink K, Cohrssen A (2004). Evaluation of vaginal complaints. *Jama* **291:** 1368-1379.

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR *et al* (2011). Enterotypes of the human gut microbiome. *Nature* **473:** 174-180.

Audirac-Chalifour A, Torres-Poveda K, Bahena-Román M, Téllez-Sosa J, Martínez-Barnetche J, Cortina-Ceballos B *et al* (2016). Cervical microbiome and cytokine profile at various stages of cervical cancer: a pilot study. *Plos One* **11:** e0153274.

Austin MN, Rabe LK, Srinivasan S, Fredricks DN, Wiesenfeld HC, Hillier SL (2015). Mageeibacillus indolicus gen. nov., sp. nov.: A novel bacterium isolated from the female genital tract. *Anaerobe* **32:** 37-42.

Bahl MI, Bergström A, Licht TR (2012). Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. *Fems Microbiology Letters* **329:** 193-197.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS *et al* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19:** 455-477.

Barnabas RV, Webb EL, Weiss HA, Wasserheit JN (2011). The role of co-infections in HIV epidemic trajectory and positive prevention: a systematic review and meta-analysis. *AIDS (London, England)* **25:** 1559.

Benning L, Golub ET, Anastos K, French AL, Cohen M, Gilbert D *et al* (2014). Comparison of lower genital tract microbiota in HIV-infected and uninfected women from Rwanda and the US. *Plos One* **9:** e96844.

Berry MA, White JD, Davis TW, Jain S, Johengen TH, Dick GJ *et al* (2017). Are oligotypes meaningful ecological and phylogenetic units? A case study of Microcystis in freshwater lakes. *Frontiers in microbiology* **8:** 365.

Beverly ES, Chen HY, Wang QJ, Zariffard MR, Cohen MH, Spear GT (2005). Utility of Amsel criteria, Nugent score, and quantitative PCR for Gardnerella vaginalis, Mycoplasma hominis, and Lactobacillus spp. for diagnosis of bacterial vaginosis in

human immunodeficiency virus-infected women. *Journal of Clinical Microbiology* **43:** 4607-4612.

Biesbroek G, Sanders EAM, Roeselers G, Wang X, Caspers MPM, Trzcinski K *et al* (2012). Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection. *Plos One* **7:** e32942.

Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R *et al* (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature methods* **10:** 57-59.

Borgdorff H, Tsivtsivadze E, Verhelst R, Marzorati M, Jurriaans S, Ndayisaba GF *et al* (2014). Lactobacillus-dominated cervicovaginal microbiota associated with reduced HIV/STI prevalence and genital HIV viral load in African women. *The ISME journal* **8:** 1781.

Borgdorff H, Armstrong SD, Tytgat HL, Xia D, Ndayisaba GF, Wastling JM *et al* (2016a). Unique insights in the cervicovaginal Lactobacillus iners and L. crispatus proteomes and their associations with microbiota dysbiosis. *Plos One* **11:** e0150767.

Borgdorff H, Gautam R, Armstrong SD, Xia D, Ndayisaba GF, van Teijlingen NH *et al* (2016b). Cervicovaginal microbiome dysbiosis is associated with proteome changes related to alterations of the cervicovaginal mucosal barrier. *Mucosal immunology* **9:** 621.

Borgdorff H, van der Veer C, van Houdt R, Alberts CJ, de Vries HJ, Bruisten SM *et al* (2017). The association between ethnicity and vaginal microbiota composition in Amsterdam, the Netherlands. *Plos One* **12:** e0181135.

Botha M, Richter K (2015). Cervical cancer prevention in South Africa: HPV vaccination and screening both essential to achieve and maintain a reduction in incidence. *SAMJ: South African Medical Journal* **105:** 33-35.

Bowman JP (1998). Pseudoalteromonas prydzensis sp. nov., a psychrotrophic, halotolerant bacterium from Antarctic sea ice. *International Journal of Systematic and Evolutionary Microbiology* **48:** 1037-1041.

Brenner DJ, Krieg NR, Staley JT (2005). *Bergey's manual of systematic bacteriology volume II: The Proteobacteria; Part B: the Gammaproteobacteria*, second edn. Editor-in-chief: GM Garrity. Springer: New York.

Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG *et al* (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology* **15:** 66.

Brotman RM, Bradford LL, Conrad M, Gajer P, Ault K, Peralta L *et al* (2012). Association between *Trichomonas vaginalis* and vaginal bacterial community composition among reproductive-age women. *Sexually transmitted diseases* **39:** 807-812.

Brotman RM, Shardell MD, Gajer P, Tracy JK, Zenilman JM, Ravel J *et al* (2014). Interplay between the temporal dynamics of the vaginal microbiota and human papillomavirus detection. *Journal of Infectious Diseases* **210:** 1723-1733.

Bruni L, Diaz M, Castellsagué M, Ferrer E, Bosch FX, de Sanjosé S (2010). Cervical human papillomavirus prevalence in 5 continents: meta-analysis of 1 million women with normal cytological findings. *Journal of Infectious Diseases* **202:** 1789-1799.

Burk RD, Harari A, Chen Z (2013). Human papillomavirus genome variants. *Virology* **445:** 232-243.

Burton JP, Reid G (2002). Evaluation of the bacterial vaginal flora of 20 postmenopausal women by direct (Nugent score) and molecular (polymerase chain reaction and denaturing gradient gel electrophoresis) techniques. *Journal of Infectious Diseases* **186:** 1770-1780.

Burton JP, Cadieux PA, Reid G (2003). Improved understanding of the bacterial vaginal microbiota of women before and after probiotic instillation. *Applied and Environmental Microbiology* **69:** 97-101.

Burton JP, Devillard E, Cadieux PA, Hammond J-A, Reid G (2004). Detection of Atopobium vaginae in postmenopausal women by cultivation-independent methods warrants further investigation. *Journal of Clinical Microbiology* **42:** 1829-1831.

Bustreo F, Okwo-Bele J, Kamara L (2015). World Health Organization perspectives on the contribution of the Global Alliance for Vaccines and Immunization on reducing child mortality. *Archives of disease in childhood* **100:** S34-S37.

Bzhalava D, Eklund C, Dillner J (2015). International standardization and classification of human papillomavirus types. *Virology* **476:** 341-344.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods* **13:** 581-583.

Callahan BJ, McMurdie PJ, Holmes SP (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal* **11:** 2639.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7:** 335-336.

Carroll IM, Ringel-Kulka T, Siddle JP, Klaenhammer TR, Ringel Y (2012). Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. *Plos One* **7:** e46953.

Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S (2007). Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology* **73:** 278-288.

Chaban B, Links MG, Jayaprakash TP, Wagner EC, Bourque DK, Lohn Z *et al* (2014). Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle. *Microbiome* **2:** 23.

Chappell CA, Rohan LC, Moncla BJ, Wang L, Meyn LA, Bunge K *et al* (2014). The effects of reproductive hormones on the physical properties of cervicovaginal fluid. *American journal of obstetrics and gynecology* **211:** 226-e221.

Chehoud C, Stieh DJ, Bailey AG, Laughlin AL, Allen SA, McCotter KL *et al* (2017). Associations of the vaginal microbiota with HIV infection, bacterial vaginosis, and demographic factors. *Aids* **31:** 895-904.

Cho I, Blaser MJ (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* **13:** 260-270.

Choi S, Kim Y, Han K, You S, Oh S, Kim S (2006). Effects of Lactobacillus strains on cancer cell proliferation and oxidative stress in vitro. *Letters in applied microbiology* **42:** 452-458.

Chow J, Tang H, Mazmanian SK (2011). Pathobionts of the gastrointestinal microbiota and inflammatory disease. *Current opinion in immunology* **23:** 473-480.

Clifford G, Smith J, Plummer M, Munoz N, Franceschi S (2003). Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis. *British journal of cancer* **88:** 63-73.

Clifford GM, Goncalves MAG, Franceschi S, HPV, group Hs (2006). Human papillomavirus types among women infected with HIV: a meta-analysis. *Aids* **20:** 2337-2344.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y *et al* (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic acids research***:** D633-D642.

Cone RA (2014). Vaginal microbiota and sexually transmitted infections that may influence transmission of cell-associated HIV. *Journal of Infectious Diseases* **210:** S616-S621.

Coolen MJ, Post E, Davis CC, Forney LJ (2005). Characterization of microbial communities found in the human vagina by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. *Applied and Environmental Microbiology* **71:** 8729-8737.

Cornejo OE, Hickey RJ, Suzuki H, Forney LJ (2018). Focusing the diversity of Gardnerella vaginalis through the lens of ecotypes. *Evolutionary Applications* **11:** 285–379.

Cundell AM (2016). Microbial ecology of the human skin. *Microbial ecology***:** 1-8.

Cuzick J, Clavel C, Petry KU, Meijer CJ, Hoyer H, Ratnam S *et al* (2006). Overview of the European and North American studies on HPV testing in primary cervical cancer screening. *International Journal of Cancer* **119:** 1095-1101.

D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC *et al* (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17:** 1.

Da Silveira MG, San Romao MV, Loureiro-Dias MC, Rombouts FM, Abee T (2002). Flow cytometric assessment of membrane integrity of ethanol-stressed Oenococcus oeni cells. *Applied and Environmental Microbiology* **68:** 6087-6093.

Damelin LH, Paximadis M, Mavri-Damelin D, Birkhead M, Lewis DA, Tiemessen CT (2011). Identification of predominant culturable vaginal Lactobacillus species and associated bacteriophages from women with and without vaginal discharge syndrome in South Africa. *Journal of medical microbiology* **60:** 180-183.

Dareng E, Ma B, Famooto A, Akarolo-Anthony S, Offiong R, Olaniyan O *et al* (2016). Prevalent high-risk HPV infection and vaginal microbiota in Nigerian women. *Epidemiology & Infection* **144:** 123-137.

Datcu R, Gesink D, Mulvad G, Montgomery-Andersen R, Rink E, Koch A *et al* (2013). Vaginal microbiome in women from Greenland assessed by microscopy and quantitative PCR. *BMC infectious diseases* **13:** 480.

Davis KM, Weiser JN (2011). Modifications to the peptidoglycan backbone help bacteria to establish infection. *Infection and immunity* **79:** 562-570.

De Vet HC, Koudstaal J, Kwee W-S, Willebrand D, Arends JW (1995). Efforts to improve interobserver agreement in histopathological grading. *Journal of clinical epidemiology* **48:** 869-873.

Demba E, Morison L, Van der Loeff MS, Awasana AA, Gooding E, Bailey R *et al* (2005). Bacterial vaginosis, vaginal flora patterns and vaginal hygiene practices in patients presenting with vaginal discharge syndrome in The Gambia, West Africa. *BMC infectious diseases* **5:** 12.

Denny L (2009). Human papillomavirus infections: Epidemiology, clinical aspects and vaccines. *The Open Infectious Diseases Journal* **3:** 135-142.

Denny L, Hendricks B, Gordon C, Thomas F, Hezareh M, Dobbelaere K *et al* (2013). Safety and immunogenicity of the HPV-16/18 AS04-adjuvanted vaccine in HIV-positive women in South Africa: a partially-blind randomised placebo-controlled study. *Vaccine* **31:** 5745-5753.

Denny L, Adewole I, Anorlu R, Dreyer G, Moodley M, Smith T *et al* (2014). Human papillomavirus prevalence and type distribution in invasive cervical cancer in sub-Saharan Africa. *International Journal of Cancer* **134:** 1389-1398.

Denny LA, Franceschi S, de Sanjosé S, Heard I, Moscicki AB, Palefsky J (2012). Human papillomavirus, human immunodeficiency virus and immunosuppression. *Vaccine* **30:** F168-F174.

Denslow SA, Westreich DJ, Firnhaber C, Michelow P, Williams S, Smith JS (2011). Bacterial vaginosis as a risk factor for high-grade cervical lesions and cancer in HIV-seropositive women. *International Journal of Gynecology & Obstetrics* **114:** 273-277.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72:** 5069-5072.

DeVos P, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA *et al* (2009). *Bergey's manual of systematic bacteriology volume III: The Firmicutes*, second edn. Managing editor: AC Parte. Springer: New York.

Di Paola M, Sani C, Clemente AM, Iossa A, Perissi E, Castronovo G *et al* (2017). Characterization of cervico-vaginal microbiota in women developing persistent high-risk Human Papillomavirus infection. *Scientific Reports* **7:** 10200.

Djigma F, Ouedraogo C, Karou D, Sagna T, Bisseye C, Zeba M *et al* (2011). Prevalence and genotype characterization of human papillomaviruses among HIV-seropositive in Ouagadougou, Burkina Faso. *Acta tropica* **117:** 202-206.

Döderlein A (1897). *Das Scheidensekret und seine Bedeutung für das Puerperalfieber*. Besold: Leipzig.

Dols JA, Smit PW, Kort R, Reid G, Schuren FH, Tempelman H *et al* (2011). Microarray-based identification of clinically relevant vaginal bacteria in relation to bacterial vaginosis. *American journal of obstetrics and gynecology* **204:** 305-e301.

Dols JA, Reid G, Kort R, Schuren FH, Tempelman H, Bontekoe T *et al* (2012). PCR-based identification of eight lactobacillus species and 18 hr-HPV genotypes in fixed cervical samples of south african women at risk of HIV and BV. *Diagnostic cytopathology* **40:** 472-477.

Donders GG, Bellen G, Grinceviciene S, Ruban K, Vieira-Baptista P (2017). Aerobic vaginitis: no longer a stranger. *Research in microbiology* **168:** 845-858.

Donders GGG, Vereecken A, Bosmans E, Dekeersmaecker A, Salembier G, Spitz B (2002). Definition of a type of abnormal vaginal flora that is distinct from bacterial vaginosis: aerobic vaginitis. *BJOG: An International Journal of Obstetrics & Gynaecology* **109:** 34-43.

Doorbar J, Quint W, Banks L, Bravo IG, Stoler M, Broker TR *et al* (2012). The biology and life-cycle of human papillomaviruses. *Vaccine* **30:** F55-F70.

Dourado MN, Aparecida Camargo Neves A, Santos DS, Araújo WL (2015). Biotechnological and agronomic potential of endophytic pink-pigmented methylotrophic Methylobacterium spp. *BioMed research international* **2015:** 909016.

Drell T, Lillsaar T, Tummeleht L, Simm J, Aaspõllu A, Väin E *et al* (2013). Characterization of the vaginal micro-and mycobiome in asymptomatic reproductive-age Estonian women. *Plos One* **8:** e54379.

Drolet M, Bénard É, Boily M-C, Ali H, Baandrup L, Bauer H *et al* (2015). Population-level impact and herd effects following human papillomavirus vaccination programmes: a systematic review and meta-analysis. *The Lancet Infectious Diseases* **15:** 565-580.

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460-2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194-2200.

El Aila NA, Tency I, Claeys G, Verstraelen H, Saerens B, dos Santos Santiago GL *et al* (2009). Identification and genotyping of bacteria from paired vaginal and rectal samples from pregnant women indicates similarity between vaginal and rectal microflora. *BMC infectious diseases* **9:** 167.

Engberts MK, Verbruggen BS, Boon ME, Van Haaften M, Heintz APM (2007). Candida and dysbacteriosis: A cytologic, population-based study of 100,605 asymptomatic women concerning cervical carcinogenesis. *Cancer Cytopathology* **111:** 269-274.

Eren AM, Zozaya M, Taylor CM, Dowd SE, Martin DH, Ferris MJ (2011). Exploring the diversity of Gardnerella vaginalis in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. *Plos One* **6:** e26732.

Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG *et al* (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution* **4:** 1111-1119.

Eren AM, Sogin ML, Maignien L (2016). New insights into microbial ecology through subtle nucleotide variation. *Frontiers in microbiology* **7:** 1318.

Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM *et al* (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* **2:** 1-7.

Faircloth BC, Glenn TC Faircloth-lab serapure protocol. 2014. http://protocols-serapure.readthedocs.org/en/latest/.

Feeney A, Sleator RD (2012). The human gut microbiome: the ghost in the machine. *Future Microbiology* **7:** 1235-1237.

Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C *et al* (2013). GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC

CancerBase No. 11 [Internet]. Available at: http://globocan.iarc.fr [Accessed April 2014].

Fettweis JM, Serrano MG, Sheth NU, Mayer CM, Glascock AL, Brooks JP *et al* (2012). Species-level classification of the vaginal microbiome. *BMC Genomics* **13:** S17.

Fettweis JM, Brooks JP, Serrano MG, Sheth NU, Girerd PH, Edwards DJ *et al* (2014). Differences in vaginal microbiome in African American women versus women of European ancestry. *Microbiology* **160:** 2272-2282.

Forney LJ, Gajer P, Williams CJ, Schneider GM, Koenig SSK, McCulle SL *et al* (2010). Comparison of self-collected and physician-collected vaginal swabs for microbiome analysis. *Journal of Clinical Microbiology* **48:** 1741-1748.

Forsum U, Jakobsson T, Larsson P, Schmidt H, Beverly A, Bjørnerem A *et al* (2002). An international study of the interobserver variation between interpretations of vaginal smear criteria of bacterial vaginosis. *Apmis* **110:** 811-818.

Fouhy F, Deane J, Rea MC, O'Sullivan Ó, Ross RP, O'Callaghan G *et al* (2015). The effects of freezing on faecal microbiota as determined using MiSeq sequencing and culture-based investigations. *Plos One* **10:** e0119355.

Frank DN, Manigart O, Leroy V, Meda N, Valéa D, Zhang W *et al* (2012). Altered vaginal microbiota are associated with perinatal mother-to-child transmission of HIV in African women from Burkina Faso. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **60:** 299-306.

Franks AH, Harmsen HJ, Raangs GC, Jansen GJ, Schut F, Welling GW (1998). Variations of bacterial populations in human feces measured by fluorescent in situ hybridization with group-specific 16S rRNA-targeted oligonucleotide probes. *Applied and Environmental Microbiology* **64:** 3336-3345.

Fredricks DN, Fiedler TL, Marrazzo JM (2005). Molecular identification of bacteria associated with bacterial vaginosis. *New England Journal of Medicine* **353:** 1899-1911.

Frisch M, Biggar RJ, Goedert JJ (2000). Human papillomavirus-associated cancers in patients with human immunodeficiency virus infection and acquired immunodeficiency syndrome. *Journal of the National Cancer Institute* **92:** 1500-1510.

Fu L, Niu B, Zhu Z, Wu S, Li W (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28:** 3150-3152.

Gagnaire A, Nadel B, Raoult D, Neefjes J, Gorvel J-P (2017). Collateral damage: insights into bacterial mechanisms that predispose host cells to cancer. *Nature Reviews Microbiology* **15:** 109.

Gajer P, Brotman RM, Bai G, Sakamoto J, Schütte UM, Zhong X *et al* (2012). Temporal dynamics of the human vaginal microbiota. *Science translational medicine* **4:** 132ra152.

Gao W, Weng J, Gao Y, Chen X (2013). Comparison of the vaginal microbiota diversity of women with and without human papillomavirus infection: a cross-sectional study. *BMC infectious diseases* **13:** 271.

Gautam R, Borgdorff H, Jespers V, Francis SC, Verhelst R, Mwaura M *et al* (2015). Correlates of the molecular vaginal microbiota composition of African women. *BMC infectious diseases* **15:** 86.

Gelber SE, Aguilar JL, Lewis KL, Ratner AJ (2008). Functional and phylogenetic characterization of Vaginolysin, the human-specific cytolysin from Gardnerella vaginalis. *Journal of Bacteriology* **190:** 3896-3903.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ *et al* (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology* **3:** 733-739.

Ghartey JP, Smith BC, Chen Z, Buckley N, Lo Y, Ratner AJ *et al* (2014). Lactobacillus crispatus dominant vaginal microbiome is associated with inhibitory activity of female genital tract secretions against Escherichia coli. *Plos One* **9:** e96659.

Gill C, van de Wijgert JH, Blow F, Darby AC (2016). Evaluation of Lysis Methods for the Extraction of Bacterial DNA for Analysis of the Vaginal Microbiota. *Plos One* **11:** e0163148.

Gillet E, Meys JF, Verstraelen H, Bosire C, De Sutter P, Temmerman M *et al* (2011). Bacterial vaginosis is associated with uterine cervical human papillomavirus infection: a meta-analysis. *BMC infectious diseases* **11:** 10.

Gillet E, Meys JF, Verstraelen H, Verhelst R, De Sutter P, Temmerman M *et al* (2012). Association between bacterial vaginosis and cervical intraepithelial neoplasia: systematic review and meta-analysis. *Plos One* **7:** e45201.

Goedert JJ, Coté TR, Virgo P, Scoppa SM, Kingma DW, Gail MH *et al* (1998). Spectrum of AIDS-associated malignant disorders. *The Lancet* **351:** 1833-1839.

Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A *et al* (2016). Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature* **201:** 6.

Gosmann C, Handley SA, Farcasanu M, Abu-Ali G, Bowman BA, Padavattan N *et al* (2017). Lactobacillus-deficient cervicovaginal bacterial communities are associated with increased HIV acquisition in young South African women. *Immunity* **46:** 29-37.

Guo F, Zhang T (2013). Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. *Applied microbiology and biotechnology* **97:** 4607-4616.

Haggerty CL, Hillier SL, Bass DC, Ness RB (2004). Bacterial vaginosis and anaerobic bacteria are associated with endometritis. *Clinical Infectious Diseases* **39:** 990-995.

Haggerty CL, Totten PA, Ferris M, Martin DH, Hoferka S, Astete SG *et al* (2009). Clinical characteristics of bacterial vaginosis among women testing positive for fastidious bacteria. *Sexually transmitted infections* **85:** 242-248.

Hale VL, Tan CL, Knight R, Amato KR (2015). Effect of preservation method on spider monkey (Ateles geoffroyi) fecal microbiota over 8 weeks. *Journal of microbiological methods* **113:** 16-26.

He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG *et al* (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature methods* **7:** 807-812.

He Y, Zhou BJ, Deng GH, Jiang XT, Zhang H, Zhou HW (2013). Comparison of microbial diversity determined with the same variable tag sequence extracted from two different PCR amplicons. *BMC Microbiol* **13:** 208.

He Y, Caporaso JG, Jiang X-T, Sheng H-F, Huse SM, Rideout JR *et al* (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* **3:** 20.

Hedge SR, Barrientes F, Desmond RA, Schwebke JR (2006). Local and systemic cytokine levels in relation to changes in vaginal flora. *Journal of Infectious Diseases* **193:** 556-562.

Hemme CL, Tu Q, Shi Z, Qin Y, Gao W, Deng Y *et al* (2015). Comparative metagenomics reveals impact of contaminants on groundwater microbiomes. *Frontiers in microbiology* **6:** 1205.

Hernández-Rodríguez C, Romero-González R, Albani-Campanario M, Figueroa-Damián R, Meraz-Cruz N, Hernández-Guerrero C (2011). Vaginal microbiota of healthy pregnant Mexican women is constituted by four Lactobacillus species and several vaginosis-associated bacteria. *Infectious diseases in obstetrics and gynecology* **2011:** 851485.

Hickey R, Abdo Z, Zhou X, Nemeth K, Hansmann M, Osborn T *et al* (2013). Effects of tampons and menses on the composition and diversity of vaginal microbial communities over time. *BJOG: An International Journal of Obstetrics & Gynaecology* **120:** 695-706.

Hiller T, Poppelreuther S, Stubenrauch F, Iftner T (2006). Comparative analysis of 19 genital human papillomavirus types with regard to p53 degradation, immortalization, phylogeny, and epidemiologic risk classification. *Cancer Epidemiology Biomarkers & Prevention* **15:** 1262-1267.

Hillier SL, Nugent RP, Eschenbach DA, Krohn MA, Gibbs RS, Martin DH *et al* (1995). Association between bacterial vaginosis and preterm delivery of a low-birth-weight infant. *New England Journal of Medicine* **333:** 1737-1742.

Hong S, Bunge J, Leslin C, Jeon S, Epstein SS (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME journal* **3:** 1365.

Hooper LV, Gordon JI (2001). Commensal host-bacterial relationships in the gut. *Science* **292:** 1115-1118.

Hoppenot C, Stampler K, Dunton C (2012). Cervical cancer screening in high-and low-resource countries: implications and new developments. *Obstetrical & gynecological survey* **67:** 658-667.

Huang Y-E, Wang Y, He Y, Ji Y, Wang L-P, Sheng H-F *et al* (2015). Homogeneity of the vaginal microbiome at the cervix, posterior fornix, and vaginal canal in pregnant Chinese women. *Microbial ecology* **69:** 407-414.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486:** 207-214.

Hummelen R, Fernandes AD, Macklaim JM, Dickson RJ, Changalucha J, Gloor GB *et al* (2010). Deep sequencing of the vaginal microbiota of women with HIV. *Plos One* **5:** e12078.

IARC (2012). Biological Agents: Human papillomaviruses. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* **100B:** 255-313.

IARC (2018). Agents classified by the IARC Monographs, Volumes 1–120 [internet] Available at: http://monographs.iarc.fr/ENG/Classification/latest_classif.php [Accessed Feb 2018]. World Health Organization.

Impey L, Child T (2012). *Obstetrics and Gynaecology*. Wiley-Blackwell.

Jarde A, Lewis-Mikhael AM, Moayyedi P, Stearns JC, Collins SM, Beyene J *et al* (2018). Pregnancy outcomes in women taking probiotics or prebiotics: a systematic review and meta-analysis. *BMC Pregnancy Childbirth* **18:** 14.

Jayaram A, Witkin SS, Zhou X, Brown CJ, Rey GE, Linhares IM *et al* (2014). The bacterial microbiome in paired vaginal and vestibular samples from women with vulvar vestibulitis syndrome. *Pathogens and disease* **72:** 161-166.

Jellouli K, Bougatef A, Manni L, Agrebi R, Siala R, Younes I *et al* (2009). Molecular and biochemical characterization of an extracellular serine-protease from *Vibrio metschnikovii* J1. *Journal of industrial microbiology & biotechnology* **36:** 939-948.

Jeon Y-S, Park S-C, Lim J, Chun J, Kim B-S (2015). Improved pipeline for reducing erroneous identification by 16S rRNA sequences using the Illumina MiSeq platform. *Journal of Microbiology* **53:** 60-69.

Jespers V, van de Wijgert J, Cools P, Verhelst R, Verstraelen H, Delany-Moretlwe S *et al* (2015). The significance of *Lactobacillus crispatus* and *L. vaginalis* for vaginal health and the negative effect of recent sex: a cross-sectional descriptive study across groups of African women. *BMC infectious diseases* **15:** 115.

Jespers V, Kyongo J, Joseph S, Hardy L, Cools P, Crucitti T *et al* (2017). A longitudinal analysis of the vaginal microbiota and vaginal immune mediators in women from sub-Saharan Africa. *Scientific Reports* **7:** 11974.

Kanagawa T (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of bioscience and bioengineering* **96:** 317-323.

Kelly HA, Sawadogo B, Chikandiwa A, Segondy M, Gilham C, Lompo O *et al* (2017). Epidemiology of high-risk human papillomavirus and cervical lesions in African women living with HIV/AIDS: effect of anti-retroviral therapy. *Aids* **31:** 273-285.

Kembel SW, Wu M, Eisen JA, Green JL (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS computational biology* **8:** e1002743.

Kennedy K, Hall MW, Lynch MDJ, Moreno-Hagelsieb G, Neufeld JD (2014). Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. *Applied and Environmental Microbiology* **80:** 5717-5722.

Kim S-N, Lee WM, Park KS, Kim JB, Han DJ, Bae J (2015). The effect of *Lactobacillus casei* extract on cervical cancer cell lines. *Contemporary Oncology* **19:** 306.

Kim TK, Thomas SM, Ho M, Sharma S, Reich CI, Frank JA *et al* (2009). Heterogeneity of vaginal microbial communities within individuals. *Journal of Clinical Microbiology* **47:** 1181-1189.

King CC, Jamieson DJ, Wiener J, Cu-Uvin S, Klein RS, Rompalo AM *et al* (2011). Bacterial vaginosis and the natural history of human papillomavirus. *Infectious diseases in obstetrics and gynecology* **2011:** 319460.

Kirby T (2015). FDA approves new upgraded Gardasil 9. *Lancet Oncology* **2:** e56.

Klebanoff MA, Schwebke JR, Zhang J, Nansel TR, Yu K-F, Andrews WW (2004). Vulvovaginal symptoms in women with bacterial vaginosis. *Obstetrics & Gynecology* **104:** 267-272.

Klomp JM, Verbruggen BSM, Korporaal H, Boon ME, de Jong P, Kramer GC *et al* (2008). *Gardnerella vaginalis* and *Lactobacillus* sp in liquid-based cervical samples in healthy and disturbed vaginal flora using cultivation-independent methods. *Diagnostic cytopathology* **36:** 277-284.

Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Prieme A, Aarestrup FM *et al* (2016). Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. *mSystems* **1:** e00095-00016.

Koeppel AF, Wu M (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic acids research* **41:** 5175-5188.

Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahé F, He Y *et al* (2016). Open-source sequence clustering methods improve the state of the art. *mSystems* **1:** e00003-00015.

Koumans EH, Sternberg M, Bruce C, McQuillan G, Kendrick J, Sutton M *et al* (2007). The prevalence of bacterial vaginosis in the United States, 2001–2004; associations with symptoms, sexual behaviors, and reproductive health. *Sexually transmitted diseases* **34:** 864-869.

Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL *et al* (2010). *Bergey's manual of systematic bacteriology volume IV: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*, second edn. Managing editor: AC Parte. Springer: New York.

Lagier J-C, Elkarkouri K, Rivet R, Couderc C, Raoult D, Fournier P-E (2013). Non contiguous-finished genome sequence and description of *Senegalemassilia anaerobia* gen. nov., sp. nov. *Standards in genomic sciences* **7:** 343.

Larkin MA, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H *et al* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947-2948.

Laubscher N, Dreyer G, Snyman L, Botha M, Van der Merwe F, Visser C *et al* (2015). The Vaccine and Cervical Cancer Screen project: experiences from a primary school-based vaccine implementation studying Gauteng and the Western Cape, South Africa. *Professional Nursing Today* **19:** 28-31.

Lederberg J (2001). Beyond the Genome. *Brooklyn Law Review*. pp 7-12.

Lederberg J, McCray A (2001). 'Ome Sweet 'Omics--A Genealogical Treasury of Words. *The Scientist* **15:** 8.

Lee CK, Herbold CW, Polson SW, Wommack KE, Williamson SJ, McDonald IR *et al* (2012). Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *Plos One* **7:** e44224.

Lee JE, Lee S, Lee H, Song Y-M, Lee K, Han MJ *et al* (2013). Association of the vaginal microbiota with human papillomavirus infection in a Korean twin cohort. *Plos One* **8:** e63514.

Leemans CR, Braakhuis BJ, Brakenhoff RH (2011). The molecular biology of head and neck cancer. *Nature Reviews Cancer* **11:** 9-22.

Lehoux M, D'Abramo CM, Archambault J (2008). Molecular mechanisms of human papillomavirus-induced carcinogenesis. *Public health genomics* **12:** 268-280.

Lehtovirta P, Paavonen J, Heikinheimo O (2008). Risk factors, diagnosis and prognosis of cervical intraepithelial neoplasia among HIV-infected women. *International journal of STD & AIDS* **19:** 37-41.

Li X, Wang H, Du X, Yu W, Jiang J, Geng Y *et al* (2017). Lactobacilli inhibit cervical cancer cell migration in vitro and reduce tumor burden in vivo through upregulation of E-cadherin. *Oncology reports* **38:** 1561-1568.

Ling Z, Kong J, Liu F, Zhu H, Chen X, Wang Y *et al* (2010). Molecular analysis of the diversity of vaginal microbiota associated with bacterial vaginosis. *BMC Genomics* **11:** 488.

Ling Z, Liu X, Luo Y, Wu X, Yuan L, Tong X *et al* (2013). Associations between vaginal pathogenic community and bacterial vaginosis in Chinese reproductive age women. *Plos One* **8:** e76589.

Linhares IM, Summers PR, Larsen B, Giraldo PC, Witkin SS (2011). Contemporary perspectives on vaginal pH and lactobacilli. *American journal of obstetrics and gynecology* **204:** 120.e121-125.

Liu SH, Cummings DA, Zenilman JM, Gravitt PE, Brotman RM (2014). Characterizing the temporal dynamics of human papillomavirus DNA detectability using short-interval sampling. *Cancer Epidemiol Biomarkers Prev* **23:** 200-208.

Ljungh Å, Wadström T (2009). *Lactobacillus molecular biology: from genomics to probiotics*. Horizon Scientific Press.

Lonky NM, Sadeghi M, Tsadik GW, Petitti D (1999). The clinical significance of the poor correlation of cervical dysplasia and cervical malignancy with referral cytologic results. *American Journal of Obstetrics & Gynecology* **181:** 560-566.

Louis P, Hold GL, Flint HJ (2014). The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology* **12:** 661.

Low N, Chersich MF, Schmidlin K, Egger M, Francis SC, Van de Wijgert JH *et al* (2011). Intravaginal practices, bacterial vaginosis, and HIV infection in women: individual participant data meta-analysis. *PLoS medicine* **8:** e1000416.

Lynch MD, Neufeld JD (2015). Ecology and exploration of the rare biosphere. *Nature reviews Microbiology* **13:** 217.

Ma B, Forney LJ, Ravel J (2012). The vaginal microbiome: rethinking health and diseases. *Annual review of microbiology* **66:** 371.

MacIntyre DA, Chandiramani M, Lee YS, Kindinger L, Smith A, Angelopoulos N *et al* (2015). The vaginal microbiome during pregnancy and the postpartum period in a European population. *Scientific Reports* **5:** 8988.

Macklaim JM, Fernandes AD, Di Bella JM, Hammond J-A, Reid G, Gloor GB (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. *Microbiome* **1:** 12.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2:** e593.

Mahé F, Rognes T, Quince C, De Vargas C, Dunthorn M (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3:** e1420.

Mahé F (2016). Fred's Metabarcoding Pipeline [internet] Available at: https://github.com/frederic-mahe/swarm/wiki/Fred's-metabarcoding-pipeline [Accessed Jun 2017].

Martin DH (2012). The microbiota of the vagina and its influence on women's health and disease. *The American journal of the medical sciences* **343:** 2.

Martin DH, Zozaya M, Lillis R, Miller J, Ferris MJ (2012). The microbiota of the human genitourinary tract: trying to see the forest through the trees. *Transactions of the American Clinical and Climatological Association* **123:** 242.

Martin HL, Richardson BA, Nyange PM, Lavreys L, Hillier SL, Chohan B *et al* (1999). Vaginal lactobacilli, microbial flora, and risk of human immunodeficiency virus type 1 and sexually transmitted disease acquisition. *Journal of Infectious Diseases* **180:** 1863-1868.

Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* **17:** 10-12.

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012). PANDAseq: paired-end assembler for illumina sequences. *BMC bioinformatics* **13:** 31.

Mathew A, George PS (2009). Trends in incidence and mortality rates of squamous cell carcinoma and adenocarcinoma of cervix–worldwide. *Asian Pac J Cancer Prev* **10:** 645-650.

Maukonen J, Simões C, Saarela M (2012). The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. *FEMS microbiology ecology* **79:** 697-708.

Mbulaiteye SM, Biggar RJ, Goedert JJ, Engels EA (2003). Immune deficiency and risk for malignancy among persons with AIDS. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **32:** 527-533.

McDonald AC, Tergas AI, Kuhn L, Denny L, Wright Jr TC (2014). Distribution of human papillomavirus genotypes among HIV-positive and HIV-negative women in Cape Town, South Africa. *Frontiers in oncology* **4**.

McMurdie PJ, Holmes S (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *Plos One* **8:** e61217.

Mehta SD, Donovan B, Weber KM, Cohen M, Ravel J, Gajer P *et al* (2015). The vaginal microbiota over an 8- to 10-Year period in a cohort of HIV-infected and HIV-uninfected women. *Plos One* **10:** e0116894.

Mitchell C, Moreira C, Fredricks D, Paul K, Caliendo AM, Kurpewski J *et al* (2009). Detection of fastidious vaginal bacteria in women with HIV infection and bacterial vaginosis. *Infectious diseases in obstetrics and gynecology* **2009:** 236919.

Mitchell C, Balkus JE, Fredricks D, Liu C, McKernan-Mullin J, Frenkel LM *et al* (2013). Interaction between lactobacilli, bacterial vaginosis-associated bacteria, and HIV Type 1 RNA and DNA Genital shedding in US and Kenyan women. *AIDS research and human retroviruses* **29:** 13-19.

Mitra A, MacIntyre D, Lee Y, Smith A, Marchesi JR, Lehne B *et al* (2015). Cervical intraepithelial neoplasia disease progression is associated with increased vaginal microbiome diversity. *Scientific Reports* **5:** 16865.

Mitra A, MacIntyre DA, Mahajan V, Lee YS, Smith A, Marchesi JR *et al* (2017). Comparison of vaginal microbiota sampling techniques: cytobrush versus swab. *Scientific Reports* **7:** 9802.

Moscicki A-B, Ellenberg JH, Farhat S, Xu J (2004). Persistence of human papillomavirus infection in HIV-infected and-uninfected adolescent girls: risk factors and differences, by phylogenetic type. *Journal of Infectious Diseases* **190:** 37-45.

Motevaseli E, Shirzad M, Akrami SM, Mousavi A-S, Mirsalehian A, Modarressi MH (2013). Normal and tumour cervical cells respond differently to vaginal lactobacilli, independent of pH and lactate. *Journal of medical microbiology* **62:** 1065-1072.

Nanda K, McCrory DC, Myers ER, Bastian LA, Hasselblad V, Hickey JD *et al* (2000). Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review. *Annals of internal medicine* **132:** 810-819.

Nearing JT, Douglas GM, Comeau AM, Langille MG (2018). Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction methods. *PeerJ PrePrints*.

Newton IL, Roeselers G (2012). The effect of training set on the classification of honey bee gut microbiota using the Naive Bayesian Classifier. *BMC Microbiol* **12:** 221.

Nguyen N-P, Warnow T, Pop M, White B (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *npj Biofilms and Microbiomes* **2:** 16004.

Nugent RP, Krohn MA, Hillier S (1991). Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *Journal of Clinical Microbiology* **29:** 297-301.

Nunn KL, Wang Y-Y, Harit D, Humphrys MS, Ma B, Cone R *et al* (2015). Enhanced trapping of HIV-1 by human cervicovaginal mucus is associated with *Lactobacillus crispatus*-dominant microbiota. *MBio* **6:** e01084-01015.

Oakley BB, Fiedler TL, Marrazzo JM, Fredricks DN (2008). Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis. *Applied and Environmental Microbiology* **74:** 4898-4909.

Oh HY, Kim B-S, Seo S-S, Kong J-S, Lee J-K, Park S-Y *et al* (2015). The association of uterine cervical microbiota with an increased risk for cervical intraepithelial neoplasia in Korea. *Clinical Microbiology and Infection* **21:** 674.e671-679.

Ojala T, Kankainen M, Castro J, Cerca N, Edelman S, Westerlund-Wikström B *et al* (2014). Comparative genomics of *Lactobacillus crispatus* suggests novel mechanisms for the competitive exclusion of *Gardnerella vaginalis*. *BMC Genomics* **15:** 1070.

Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R *et al* (2015). Vegan: Community Ecology Package. R package version 2.3-2. 2015. http://CRAN.R-project.org/package=vegan.

Östör AG (1993). Natural history of cervical intraepithelial neoplasia: a critical review. *International Journal of Gynecological Pathology* **12:** 186.

Pendharkar S, Magopane T, Larsson P-G, de Bruyn G, Gray GE, Hammarström L *et al* (2013). Identification and characterisation of vaginal lactobacilli from South African women. *BMC infectious diseases* **13:** 43.

Pépin J, Deslandes S, Giroux G, Sobéla F, Khonde N, Diakité S *et al* (2011). The complex vaginal flora of west african women with bacterial vaginosis. *Plos One* **6:** e25082.

Pett M, Coleman N (2007). Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *The Journal of pathology* **212:** 356-367.

Pinheiro J, Bates D, DebRoy S, Sarkar D (2017). R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1-131. *Available at h ttp://CRAN R-project org/package= nlme*.

Piyathilake CJ, Ollberding NJ, Kumar R, Macaluso M, Alvarez RD, Morrow CD (2016). Cervical microbiota associated with higher grade cervical intraepithelial neoplasia in women infected with high-risk human papillomaviruses. *Cancer Prevention Research* **9:** 357-366.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research* **35:** 7188-7196.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. 2015. www.R-project.org.

Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL *et al* (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences* **108:** 4680-4687.

Redelinghuys MJ, Ehlers MM, Bezuidenhoudt J, Becker PJ, Kock MM (2017). Assessment of *Atopobium vaginae* and *Gardnerella vaginalis* concentrations in a cohort of pregnant South African women. *Sex Transm Infect* **93:** 410-415.

Reimers LL, Mehta SD, Massad LS, Burk RD, Xie X, Ravel J *et al* (2016). The cervicovaginal microbiota and its associations with human papillomavirus detection

in HIV-infected and HIV-uninfected women. *The Journal of infectious diseases* **214:** 1361-1369.

Richart RM, Barron BA (1969). A follow-up study of patients with cervical dysplasia. *American journal of obstetrics and gynecology* **105:** 386-393.

Roesch LF, Casella G, Simell O, Krischer J, Wasserfall CH, Schatz D *et al* (2009). Influence of fecal sample storage on bacterial community diversity. *The open microbiology journal* **3:** 40.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4:** e2584.

Romanenko LA, Zhukova NV, Lysenko AM, Mikhailov VV, Stackebrandt E (2003). Assignment of '*Alteromonas marinoglutinosa*'NCIMB 1770 to *Pseudoalteromonas mariniglutinosa* sp. nov., nom. rev., comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **53:** 1105-1109.

Romero R, Hassan SS, Gajer P, Tarca AL, Fadrosh DW, Nikita L *et al* (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* **2:** 4.

Rosenthal RS, Blundell JK, Perkins HR (1982). Strain-related differences in lysozyme sensitivity and extent of O-acetylation of gonococcal peptidoglycan. *Infection and immunity* **37:** 826-829.

Russo F, Orlando A, Linsalata M, Cavallini A, Messa C (2007). Effects of *Lactobacillus rhamnosus* GG on the cell growth and polyamine metabolism in HGC-27 human gastric cancer cells. *Nutrition and cancer* **59:** 106-114.

Rylev M, Bek-Thomsen M, Reinholdt J, Ennibi OK, Kilian M (2011). Microbiological and immunological characteristics of young Moroccan patients with aggressive periodontitis with and without detectable *Aggregatibacter actinomycetemcomitans* JP2 infection. *Molecular oral microbiology* **26:** 35-51.

Salonen A, Nikkilä J, Jalanka-Tuovinen J, Immonen O, Rajilić-Stojanović M, Kekkonen RA *et al* (2010). Comparative analysis of fecal DNA extraction methods

with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *Journal of microbiological methods* **81:** 127-134.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF *et al* (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12:** 87.

Sankaranarayanan R, Nessa A, Esmy PO, Dangou J-M (2012). Visual inspection methods for cervical cancer prevention. *Best practice & research Clinical obstetrics & gynaecology* **26:** 221-232.

Schaechter M (2013). [cited 15 June 2015] The Result of the Microbiome Poll [Internet]. Available from: http://schaechter.asmblog.org/schaechter/2013/12/the-result-of-the-microbiome-poll.html.

Schellenberg JJ, Links MG, Hill JE, Dumonceaux TJ, Kimani J, Jaoko W *et al* (2011). Molecular definition of vaginal microbiota in East African commercial sex workers. *Applied and Environmental Microbiology* **77:** 4066-4074.

Scheurer M, Tortolero-Luna G, Adler-Storthz K (2005). Human papillomavirus infection: biology, epidemiology, and prevention. *International Journal of Gynecological Cancer* **15:** 727-746.

Schiller JT, Day PM, Kines RC (2010). Current understanding of the mechanism of HPV infection. *Gynecologic oncology* **118:** S12-S17.

Schindler CA, Schuhardt VT (1964). Lysostaphin: a new bacteriolytic agent for the *Staphylococcus*. *Proceedings of the National Academy of Sciences* **51:** 414-421.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic acids research* **43:** e37-e37.

Schultz W (1961). Bakteriologie und Lochien. *Verhandlungen der Deutschen Gesellschaft für Gynäkologie*. Springer. pp 126-137.

Schwenger EM, Tejani AM, Loewen PS (2015). Probiotics for preventing urinary tract infections in adults and children. *Cochrane Database of Systematic Reviews*.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS *et al* (2011). Metagenomic biomarker discovery and explanation. *Genome biology* **12:** R60.

Shipitsyna E, Roos A, Datcu R, Hallén A, Fredlund H, Jensen JS *et al* (2013). Composition of the vaginal microbiota in women of reproductive age – sensitive and specific molecular diagnosis of bacterial vaginosis is possible? *Plos One* **8:** e60670.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W *et al* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7:** 539.

Sievert C, Parmer C, Hocking T, Chamberlain S, Ram K, Corvellec M *et al* (2017). plotly: create interactive web graphics via “plotly. js”. R package version 4.7.1.

Simoes J, Discacciati M, Brolazo E, Portugal P, Dini D, Dantas M (2006). Clinical diagnosis of bacterial vaginosis. *International Journal of Gynecology & Obstetrics* **94:** 28-32.

Skillings JH, Mack GA (1981). On the use of a Friedman-type statistic in balanced and unbalanced block designs. *Technometrics* **23:** 171-177.

Smith BC, McAndrew T, Chen Z, Harari A, Barris DM, Viswanathan S *et al* (2012). The cervical microbiome over 7 years and a comparison of methodologies for its characterization. *Plos One* **7:** e40425.

Smith JS, Lindsay L, Hoots B, Keys J, Franceschi S, Winer R *et al* (2007). Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: A meta-analysis update. *International Journal of Cancer* **121:** 621-632.

Smith WL, Hedges SR, Mordechai E, Adelson ME, Trama JP, Gygax SE *et al* (2014). Cervical and vaginal flora specimens are highly concordant with respect to bacterial vaginosis-associated organisms and commensal lactobacillus species in women of reproductive age. *Journal of Clinical Microbiology* **52:** 3078-3081.

Solomon D, Davey D, Kurman R, Moriarty A, O'connor D, Prey M *et al* (2002). The 2001 Bethesda System: terminology for reporting results of cervical cytology. *Jama* **287:** 2114-2119.

Song SJ, Amir A, Metcalf JL, Amato KR, Xu ZZ, Humphrey G *et al* (2016). Preservation methods differ in fecal microbiome stability, affecting suitability for field studies. *mSystems* **1:** e00021-00016.

Spear GT, Sikaroodi M, Zariffard MR, Landay AL, French AL, Gillevet PM (2008). Comparison of the diversity of the vaginal microbiota in HIV-infected and HIV-uninfected women with or without bacterial vaginosis. *Journal of Infectious Diseases* **198:** 1131-1140.

Spear GT, Gilbert D, Landay AL, Zariffard R, French AL, Patel P *et al* (2011). Pyrosequencing of the genital microbiotas of HIV-seropositive and-seronegative women reveals *Lactobacillus iners* as the predominant *Lactobacillus* Species. *Applied and Environmental Microbiology* **77:** 378-381.

Spiegel C, Amsel R, Holmes K (1983). Diagnosis of bacterial vaginosis by direct gram stain of vaginal fluid. *Journal of Clinical Microbiology* **18:** 170-177.

Srinivasan S, Hoffman NG, Morgan MT, Matsen FA, Fiedler TL, Hall RW *et al* (2012). Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *Plos One* **7:** e37818.

Strickler HD, Palefsky JM, Shah KV, Anastos K, Klein RS, Minkoff H *et al* (2003). Human papillomavirus type 16 and immune status in human immunodeficiency virus-seropositive women. *Journal of the National Cancer Institute* **95:** 1062-1071.

Suzuki MT, Giovannoni SJ (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology* **62:** 625-630.

Tankovic J, Timinskas A, Janulaitiene M, Zilnyte M, Baudel J-L, Maury E *et al* (2017). *Gardnerella vaginalis* bacteremia associated with severe acute encephalopathy in a young female patient. *Anaerobe* **47:** 132-134.

Tedjo DI, Jonkers DM, Savelkoul PH, Masclee AA, van Best N, Pierik MJ *et al* (2015). The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. *Plos One* **10:** e0126685.

Tikhonov M, Leach RW, Wingreen NS (2015). Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *The ISME journal* **9:** 68-80.

Tjalma W, Van Waes T, Van den Eeden L, Bogers J (2005). Role of human papillomavirus in the carcinogenesis of squamous cell carcinoma and adenocarcinoma of the cervix. *Best practice & research Clinical obstetrics & gynaecology* **19:** 469-483.

Toft L, Tolstrup M, Storgaard M, Østergaard L, Søgaard OS (2014). Vaccination against oncogenic human papillomavirus infection in HIV-infected populations: review of current status and future perspectives. *Sexual health* **11:** 511-523.

Tommasino M (2014). *The human papillomavirus family and its role in carcinogenesis*, vol. 26. Academic Press.

Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T *et al* (2015). Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in microbiology* **6:** 771.

UNAIDS (2013). Global report: UNAIDS report on the global AIDS epidemic 2013: Geneva: UNAIDS.

Ursell LK, Gunawardana M, Chang S, Mullen M, Moss JA, Herold BC *et al* (2014). Comparison of the vaginal microbial communities in women with recurrent genital HSV receiving acyclovir intravaginal rings. *Antiviral research* **102:** 87-94.

Van de Wijgert JH, Verwijs MC, Turner AN, Morrison CS (2013). Hormonal contraception decreases bacterial vaginosis but oral contraception may increase candidiasis: implications for HIV transmission. *Aids* **27:** 2141-2153.

van de Wijgert JH, Borgdorff H, Verhelst R, Crucitti T, Francis S, Verstraelen H *et al* (2014). The vaginal microbiota: what have we learned after a decade of molecular characterization? *Plos One* **9:** e105998.

Van Den Heuvel R, Van Der Biezen E, Jetten MSM, Hefting MM, Kartal B (2010). Denitrification at pH 4 by a soil-derived *Rhodanobacter*-dominated community. *Environmental microbiology* **12:** 3264-3271.

Van Essen HF, Verdaasdonk MA, Elshof SM, de Weger RA, van Diest PJ (2010). Alcohol based tissue fixation as an alternative for formaldehyde: influence on immunohistochemistry. *Journal of clinical pathology***:** jcp. 2010.079905.

Venables WN, Ripley BD (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Verstraelen H, Verhelst R, Claeys G, De Backer E, Temmerman M, Vaneechoutte M (2009). Longitudinal analysis of the vaginal microflora in pregnancy suggests that *L. crispatus* promotes the stability of the normal vaginal microflora and that *L. gasseri* and/or *L. iners* are more conducive to the occurrence of abnormal vaginal microflora. *BMC Microbiology* **9:** 116.

Vilos GA (1998). The history of the Papanicolaou smear and the odyssey of George and Andromache Papanicolaou. *Obstetrics & Gynecology* **91:** 479-483.

Virtanen S, Kalliala I, Nieminen P, Salonen A (2017). Comparative analysis of vaginal microbiota sampling using 16S rRNA gene analysis. *Plos One* **12:** e0181477.

Vlčková K, Mrázek J, Kopečný J, Petrželková KJ (2012). Evaluation of different storage methods to characterize the fecal bacterial communities of captive western lowland gorillas (*Gorilla gorilla gorilla*). *Journal of microbiological methods* **91:** 45-51.

Wagner Mackenzie B, Waite DW, Taylor MW (2015). Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. *Frontiers in microbiology* **6:** 130.

Wakeham K, Kavanagh K (2014). The burden of HPV-associated anogenital cancers. *Current oncology reports* **16:** 1-11.

Walboomers JMM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV *et al* (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *Journal of Pathology* **189:** 12-19.

Walther-António MRS, Jeraldo P, Berg Miller ME, Yeoman CJ, Nelson KE, Wilson BA *et al* (2014). Pregnancy's stronghold on the vaginal microbiome. *Plos One* **9:** e98514.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73:** 5261-5267.

Watts DH, Fazarri M, Minkoff H, Hillier SL, Sha B, Glesby M *et al* (2005). Effects of bacterial vaginosis and other genital infections on the natural history of human papillomavirus infection in HIV-1–infected and high-risk HIV-1–uninfected women. *Journal of Infectious Diseases* **191:** 1129-1139.

Weng S-L, Chiu C-M, Lin F-M, Huang W-C, Liang C, Yang T *et al* (2014). Bacterial communities in semen from men of infertile couples: metagenomic sequencing reveals relationships of seminal microbiota to semen quality. *Plos One* **9:** e110152.

Weon H-Y, Kim B-Y, Hong S-B, Jeon Y-A, Kwon S-W, Go S-J *et al* (2007). *Rhodanobacter ginsengisoli* sp. nov. and *Rhodanobacter terrae* sp. nov., isolated from soil cultivated with Korean ginseng. *International Journal of Systematic and Evolutionary Microbiology* **57:** 2810-2813.

Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG *et al* (2012). Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME Journal* **6:** 94-103.

Wertz J, Isaacs-Cosgrove N, Holzman C, Marsh TL (2009). Temporal shifts in microbial communities in nonpregnant African-American women with and without bacterial vaginosis. *Interdisciplinary perspectives on infectious diseases* **2008:** 181253.

Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R *et al* (2014). Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2:** 19.

Westcott SL, Schloss PD (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3:** e1487.

Whipps J, Lewis K, Cooke R (1988). Mycoparasitism and plant disease control. *Fungi in biological control systems***:** 161-187.

WHO (2010). Antiretroviral therapy for HIV infection in adults and adolescents: recommendations for a public health approach-2010 revision. *World Health Organisation Geneva, Switzerland*.

Wiesenfeld HC, Hillier SL, Krohn MA, Landers DV, Sweet RL (2003). Bacterial vaginosis is a strong predictor of *Neisseria gonorrhoeae* and *Chlamydia trachomatis* infection. *Clinical Infectious Diseases* **36:** 663-668.

Willner D, Daly J, Whiley D, Grimwood K, Wainwright CE, Hugenholtz P (2012). Comparison of DNA extraction methods for microbial community profiling with an application to pediatric bronchoalveolar lavage samples. *Plos One* **7:** e34605.

Wira CR, Rossoll RM, Kaushic C (2000). Antigen-presenting cells in the female reproductive tract: influence of estradiol on antigen presentation by vaginal cells. *Endocrinology* **141:** 2877-2885.

Witkin SS, Mendes-Soares H, Linhares IM, Jayaram A, Ledger WJ, Forney LJ (2013). Influence of vaginal bacteria and D-and L-lactic acid isomers on vaginal extracellular matrix metalloproteinase inducer: implications for protection against upper genital tract infections. *MBio* **4:** e00460-00413.

Wright TC (2007). Cervical cancer screening in the 21st century: is it time to retire the PAP smear? *Clinical obstetrics and gynecology* **50:** 313-323.

Xie HY, Feng D, Wei DM, Mei L, Chen H, Wang X *et al* (2017). Probiotics for vulvovaginal candidiasis in non-pregnant women. *Cochrane Database of Systematic Reviews*.

Xu M, Luo W, Elzi DJ, Grandori C, Galloway DA (2008). NFX1 interacts with mSin3A/histone deacetylase to repress hTERT transcription in keratinocytes. *Molecular and cellular biology* **28:** 4819-4828.

Yamamoto T, Zhou X, Williams CJ, Hochwalt A, Forney LJ (2009). Bacterial populations in the vaginas of healthy adolescent women. *Journal of pediatric and adolescent gynecology* **22:** 11-18.

Yang J, Summanen PH, Henning SM, Hsu M, Lam H, Huang J *et al* (2015). Xylooligosaccharide supplementation alters gut bacteria in both healthy and prediabetic adults: a pilot study. *Frontiers in physiology* **6:** 216.

Yeoman CJ, Yildirim S, Thomas SM, Durkin AS, Torralba M, Sutton G *et al* (2010). Comparative genomics of *Gardnerella vaginalis* strains reveals substantial differences in metabolic and virulence potential. *Plos One* **5:** e12411.

Yeoman CJ, Thomas SM, Miller M, Ulanov AV, Torralba M, Lucas S *et al* (2013). A multi-omic systems-based approach reveals metabolic markers of bacterial vaginosis and insight into the disease. *Plos One* **8:** e56111.

Yokogawa K, Kawata S, Nishimura S, Ikeda Y, Yoshimura Y (1974). Mutanolysin, bacteriolytic agent for cariogenic streptococci: partial purification and properties. *Antimicrobial agents and chemotherapy* **6:** 156-165.

Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA, Elshahed MS (2009). Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Applied and Environmental Microbiology* **75:** 5227-5236.

Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. *Plos One* **7:** e33865.

Zarakolu P, Hodoglugil NNS, Aydın F, Tosun I, Gozalan A, Unal S (2004). Reliability of interpretation of gram-stained vaginal smears by nugent's scoring system for diagnosis of bacterial vaginosis. *Diagnostic microbiology and infectious disease* **48:** 77-80.

Zhang J, Kobert K, Flouri T, Stamatakis A (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30:** 614-620.

Zhang Z, Schwartz S, Wagner L, Miller W (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7:** 203-214.

Zhou X, Hansmann MA, Davis CC, Suzuki H, Brown CJ, Schütte U *et al* (2010). The vaginal bacterial communities of Japanese women resemble those of women in other racial groups. *FEMS Immunology & Medical Microbiology* **58:** 169-181.

Zur Hausen H (2002). Papillomaviruses and cancer: from basic studies to clinical application. *Nature Reviews Cancer* **2:** 342-350.

Appendix A: Gel Image Sample P01

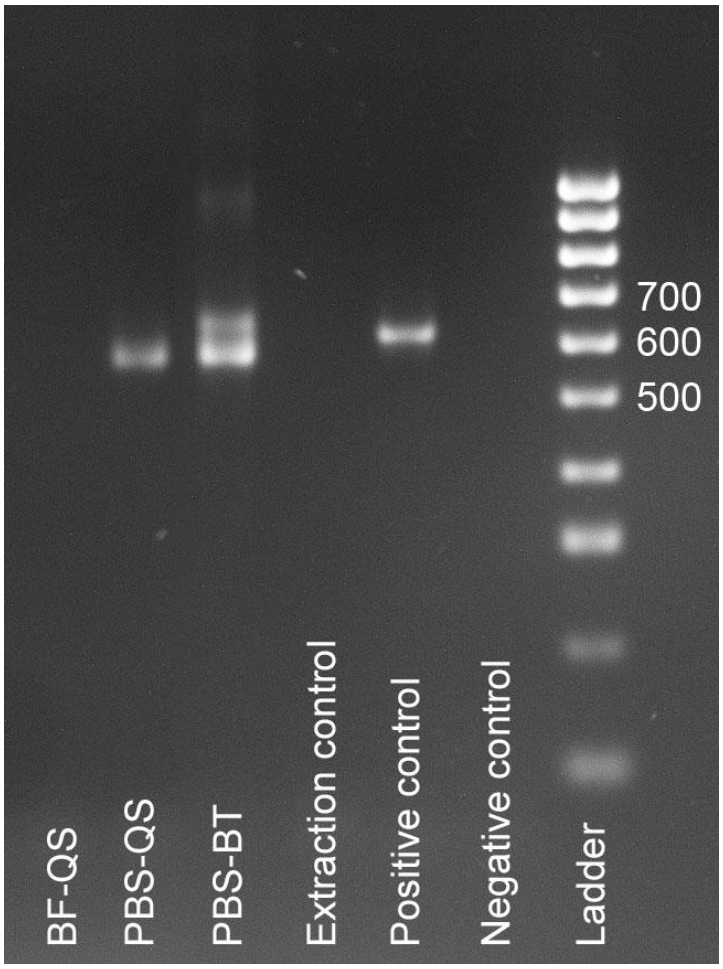


Figure A.1 Gel image of sample P01 described in section 2.3.2. The ladder is Bioline Hyperladder IV, this corresponding sizes of relevant bands are labelled in base pairs. The concentration of DNA in the BF-QS extract was too low to show a band. The PBS-QS extract shows a single band below 600 bp in size while the PBS-BT extract produced a double band around 600 bp. The positive control contains *Lactobacillus amylovorus* DNA, with an expected band size of 605 bp. The extraction control was produced from nuclease free water that was taken through the DNA extraction process alongside samples. The negative control is a PCR control that contained nuclease free water.

Appendix B: Reagent Contamination

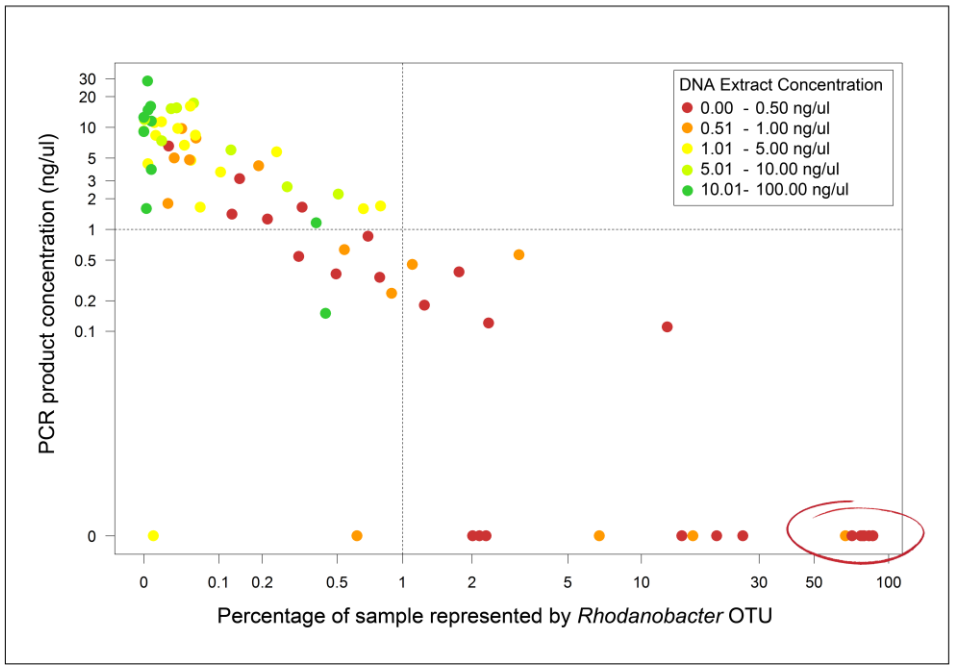


Figure B.1 Scatterplot of PCR product DNA concentration against the percentage of sample made up of *Rhodanobacter* OTU for each of the vaginal sample extracts described in section 2.2 (along with additional samples sequenced on the same run but not presented as part of this thesis). *Rhodanobacter* was the major contaminant of this sequencing run and can be thought of as a proxy for the degree of sample contamination. The DNA concentration of the sample extract (prior to PCR) is indicated by colour. Negative extraction controls are circled. Note that the axes are on a log scale. There is a significant negative correlation between PCR product concentration and the percentage of *Rhodanobacter* OTU (Spearman's rank correlation coefficient -0.85; $p < 0.0001$). There is also a significant negative correlation between DNA extract concentration and the percentage of *Rhodanobacter* OTU (Spearman's rank correlation coefficient -0.76; $p < 0.0001$). The relationship is likely to be less strong due to the fact that the relative abundance of contaminants is dependent on the amount of bacterial DNA in the sample and PCR product concentration provides a better estimation of this than DNA extract concentration (because the latter is also affected by the presence of eukaryotic DNA). In order to achieve a low level of contamination (defined arbitrarily as <1% of *Rhodanobacter*), PCR product concentrations of 1 ng/µl or higher are optimal, which can usually be reached for DNA extracts with concentrations of 1 ng/µl or higher (over 25 PCR cycles and by adding 10 µl to a 25 µl reaction). As is evident from the graph above, it was difficult to achieve this concentration with some of the samples, explaining the high levels of contamination. This is in part due to the removal of the swab head from these samples prior to analysis which is likely to have reduced yield (see section 2.4 and 2.5).

Appendix C: OTU Picking in QIIME with USEARCH

In order to compare OTU picking using the pick_otus.py script in Quantitative Insights Into Microbial Ecology (QIIME v. 1.8.0)(Caporaso et al 2010) with method "usearch" using USEARCH v. 5.2.236 (Edgar 2010) with method "usearch61" using USEARCH v. 6.1.544 (Edgar 2010), read data from the samples described in section 2.2 were trimmed using Cutadapt v. 1.2.1 (Martin 2011), error corrected using SPAdes v 3.1.0 (Bankevich et al 2012) and paired-end aligned using PEAR v0.9.6 (Zhang et al 2014). All samples were included in *de novo* OTU picking as the presence of samples may affect clustering. Apart from the choice of method described above, default settings were used. The resulting profiles of the monoculture positive control sample were then compared. While method "usearch" resulted in lower OTU number across the whole dataset compared to method "usearch61" (N=567 compared with N=80,756), the former resulted in several larger OTUs in the positive control (Figure C.1). This appears to be due to an additional error correction step applied when using method "usearch" that aims to correct read error, but appears to erraneously inflate OTUs in this control sample.

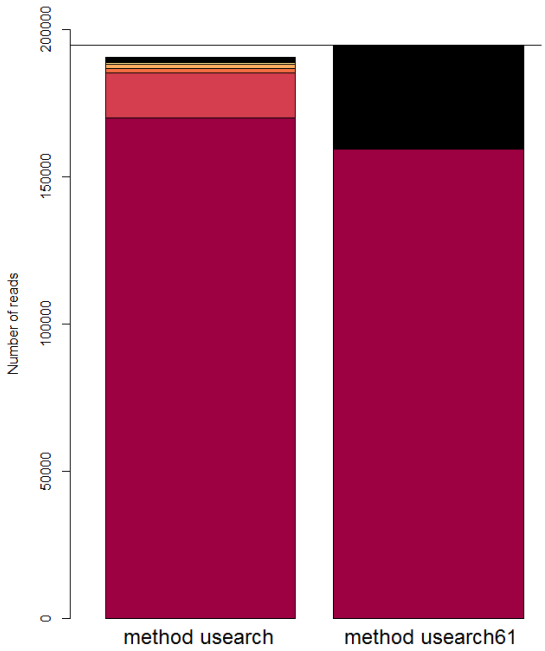


Figure C.1 Positive control (monoculture of *Lactobacillus amylovorus*) profiles created using method "usearch" and method "usearch61" using the pick_otus.py workflow script in QIIME. Horizontal line represents total number of reads aligned with PEAR. By default, method "usearch" excludes low abundance OTUs in the dataset as probable error (<4 reads in total), explaining the loss of reads with this method. Each colour represents a different OTU within each profile, but does not indicate equivalence across the two profiles. The black regions represent OTUs that are too small to resolve.

Appendix D: Base Quality and Paired-End Alignment

Including the primer region, the V3-V4 region sequenced in this study is between 440 and 474 bases in length. This means that, with an overlap region of 10 bases (the default minimum overlap using paired-end aligner PEAR), the paired end reads must be at least a combined length of 485 bases in order to cover all bacterial species that may be present in the sample. Therefore, using the Illumina 2x250bp system, if more than 15 bases in total are trimmed due to poor quality, the paired ends cannot be aligned. In the case of the two Illumina MiSeq runs described in section 2.6, the median phred score for read 1 was above 30 (=probability of incorrect base call is 1 in 1000) for both runs, indicating good base quality for read 1. For read 2, the median phred score for the first sequencing run (samples described in section 2.6.2) was 30 at read position 240-249 and 24 at read position 250 (see Figure D.1). By comparison, the median phred score for the second sequencing run (samples described in section 2.6.5) was 17 in region 240-259 and 2 at read position 250 (see Figure D.1). The quality score of base calls is expected to decline as the run progresses resulting in a gradual drop of quality towards the end of a read. Poor read quality is usually dealt with in the bioinformatics pipeline, by trimming (i.e. removing low quality bases from the end of the read) to reduce the risk of erroneous results from incorrectly called bases. In the original bioinformatics pipeline, reads were demultiplexed and trimmed for the presence of Illumina adapter sequences and low quality bases (quality threshold Q = 20) using Cutadapt v. 1.2.1 (Martin 2011) and Sickle v. 1.200 (github.com/najoshi/sickle), respectively. The resulting reads were error corrected using SPAdes v 3.1.0 (Bankevich et al 2012) and paired-end alignment was performed using PANDAseq v. 2.4 (Masella et al 2012). The obtained sequences were then binned into operational taxonomic units (OTUs) based on 97% sequence similarity using USEARCH v. 5.2.236 (Edgar 2010) through Quantitative Insights Into Microbial Ecology (QIIME v. 1.7.0)(Caporaso et al 2010). This pipeline resulted in significantly lower alignment success for samples on the second run (median alignment success rate 88% vs. 45%; $P < 0.0001$ by repeated measures ANOVA). Interestingly, alignment success appeared to be sample-dependent (i.e. all extraction methods showed the same pattern of alignment success rate with extracts from the same sample either aligning comparatively well across all methods, or comparatively poorly across all methods). The reason for this is likely to be sample composition since there was a differential effect on different OTUs, causing apparent disappearance of specific OTUs across all samples on the second sequencing run. This affected OTUs identified as

Prevotella sp. (two OTUs), *Veillonella* sp., *Atopobium* sp., *Dialister* sp., *Mycoplasma* sp., *Ureaplasma* sp., BVAB2 and *Mageeibacillus indolicus* (BVAB3). One of the *Prevotella* sp. OTUs, was absent from the BF extract (second sequencing run), but represented between 11 – 13% of all other extracts (first sequencing run). No difference in GC content of these OTUs and other OTUs could be identified (both along the length of the sequence and in the central region), nor was there any difference in sequence length. Even though it has been reported that 2x250bp compares well to 2x300bp using this 16S rRNA region (Fadrosh et al 2014) and the loss of reads could be at least partially corrected by altering the bioinformatics pipeline, it was decided that future sequencing would be performed using the 2x300bp system rather than the original 2x250bp. However, the fact that quality trimming can dramatically and differentially affects OTU presence in the results should be borne in mind, and alignment success rates should always be evaluated and poor success rates investigated. This is particularly important when combining data from different runs that may differ in quality, but also between different samples on the same run. This is further highlighted by this same effect being observed with a positive control sample on a different Illumina MiSeq sequencing run (not presented in this thesis) that contained only *Prevotella bivia* LMG6452 DNA. Whilst the alignment success of samples was variable (range 23-91%), for this control only 1% of reads aligned using the pipeline above. In light of these observations, quality checking this type of data for reads lost during the alignment step and investigating any discrepancies is recommended.

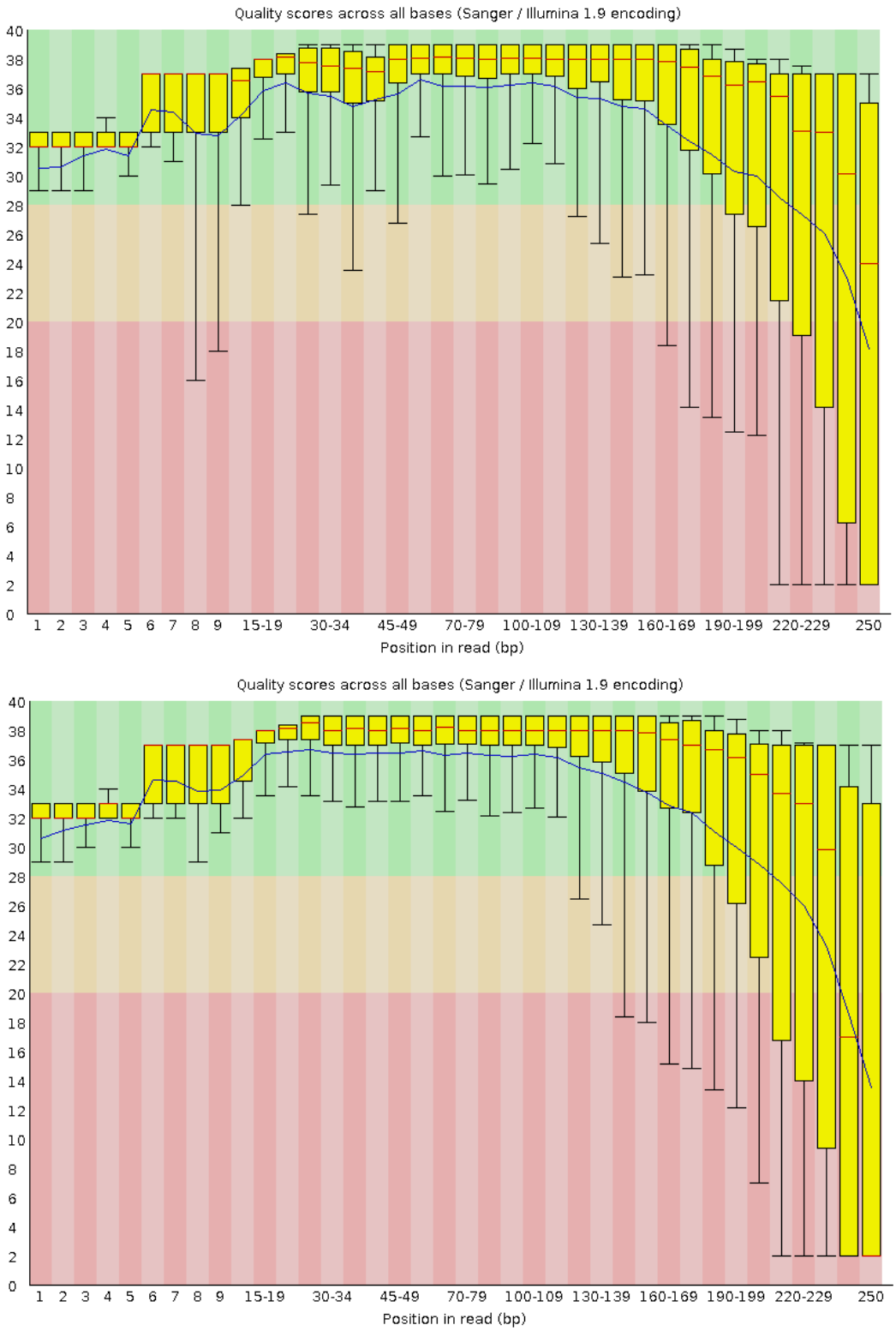


Figure D.1 Box and Whisker plot of per base sequence quality of read 2 for the two Illumina MiSeq runs described in sections 2.6.3 (upper graph) and 2.6.5 (lower graph). Boxes extend from the lower quartiles to the upper quartiles with median values indicated by the line within each box. Whiskers represent the 90th and 10th percentiles. The blue line represents the mean score. There is a significantly lower average base quality on the second run. Graphs generated by FastQC (version 0.10.1).

Appendix E: Global Similarity of Lactobacillus 16S rRNA Regions

In the vaginal environment, *Lactobacillus* species are often well represented and previous studies have shown significant differences in the health benefits afforded by different *Lactobacillus* species. The ability to differentiate between them is therefore crucial for 16S rRNA amplicon sequencing studies of the vaginal microbiome. In order to determine what effect the choice of PCR primers would have on the ability to differentiate between the different *Lactobacillus* species, we used the following approach:

Acquisition of sequence data

Sequences of the 16S rRNA gene from lactobacilli species that had been reported in vaginal samples (Ljungh and Wadström 2009) were obtained from GenBank. Sequences were available from genome assemblies of the following species: *L. acidophilus*, *L. brevis*, *L. crispatus*, *L. delbrueckii*, *L. fermentum*, *L. gasseri*, *L. helveticus*, *L. jensenii*, *L. johnsonii*, *L. paracasei*, *L. plantarum*, *L. reuteri* and *L. rhamnosus* (see Table D.1). All sequences that were part of a complete genome assembly were obtained including each copy of the gene within the same genome (also including copies that had not been annotated as 16S rRNA genes but were identified as such based on sequence). Identical sequences from the same genome assembly were then removed. Additionally, sequences were obtained from partial sequences of the 16S ribosomal RNA gene from the following species for which whole genome assemblies were not available: *L. gallinarum*, *L. iners* and *L. vaginalis* (see Table D.1). Three commonly used regions of the 16S rRNA gene were selected based on their previous use in vaginal microbiome studies: V1-V2, V3-V4 and V6. The sequences for the V1-V2 region were extracted from each of the sequences above using the primer sequences AGAGTTTGATCCTGGCTCAG and TACGGYAGGCAGCAG (Jeon et al 2015). The sequences for the V3-V4 region were extracted using the primer sequences ACTCCTACGGRAGGCAGCAG and GGATTAGATACCCTGGTAGTC (Jeon et al 2015). The sequences for the V6 region were extracted using the primer sequences ACTYAAAKGAATTGRCGGGG and GARCTGRCGRCRRCCATGCA (Smith et al 2012).

Table E.1 List of GenBank bacterial 16S sequences used in this study.

Bacterial strain	Genome assembly	Number of 16S copies identified
L. acidophilus 30SC	Y	4
L. acidophilus La-14	Y	4
L. acidophilus NCFM	Y	4
L. brevis ATCC367	Y	5
L. brevis KB290	Y	5
L. crispatus ST1	Y	4
L. delbrueckii 2038	Y	9
L. delbrueckii ATCC BAA-365	Y	9
L. delbrueckii ATCC11842	Y	9
L. delbrueckii ND02	Y	9
L. fermentum CECT5716	Y	5
L. fermentum F-6	Y	5
L. fermentum IFO3956	Y	5
L. gasseri ATCC33323	Y	6
L. helveticus CNRZ32	Y	4
L. helveticus DPC4571	Y	4
L. helveticus H10	Y	4
L. helveticus H9	Y	4
L. helveticus R0052	Y	4
L. jensenii JV-V16	Y	2
L. johnsonii DPC6026	Y	4
L. johnsonii FI9785	Y	4
L. johnsonii N6.2	Y	4
L. johnsonii NCC533	Y	6
L. paracasei 8700_2	Y	5
L. paracasei N1115	Y	5
L. plantarum 16	Y	5
L. plantarum JDM1	Y	5
L. plantarum p-8	Y	5
L. plantarum ST-III	Y	5
L. plantarum WCFS1	Y	5
L. plantarum ZJ316	Y	5
L. reuteri DSM20016	Y	6
L. reuteri I5007	Y	6
L. reuteri JCM1112	Y	6
L. reuteri SD2112	Y	6
L. reuteri TD1	Y	6
L. rhamnosus ATCC53103	Y	5
L. rhamnosus ATCC8530	Y	5
L. rhamnosus GG	Y	5
L. rhamnosus Lc705	Y	5
L. rhamnosus LOCK900	Y	5
L. rhamnosus LOCK908	Y	5
L. gallinarum ATCC33199	N	N/A
L. iners CIP109878T	N	N/A
L. vaginalis DoxG3	N	N/A

Sequence alignment and percentage identity

For each region, lactobacillus sequences were aligned using Clustal 2.1 (Larkin et al 2007) and, after the alignment was visually checked for accuracy, the associated percentage identity matrix was used to determine sequence similarity. Primer regions were not included in the alignment, resulting in sequences totalling between

335-348 bases in length for the V1-V2 region, 427 bases in length for the V3-V4 region and 123 bases in length for the V6 region.

Results

Within-species sequence similarity

For those lactobacilli for which a genome assembly was available, between 2 (*L. jensenii*) and 9 copies (*L. delbrueckii*) of the 16S rRNA gene were identified within the same genome (see Table D.1). For region V1-V2, sequence similarity within the same genome (i.e. between multiple copies of the 16S rRNA gene) ranged from 98.0-100.0%, with the lowest sequence similarity found within *L. reuteri* strain DSM20016. For region V3-V4, sequence similarity within the same genome ranged from 98.6-100.0%, with the lowest sequence similarity found within *L. plantarum* strain 16. For region V6, sequence similarity within the same genome ranged from 87.0-100.0%, with the lowest sequence similarity found within *L. jensenii* strain JV-V16. Additionally, *L. delbrueckii* strain ND02 also had a within-genome similarity below 97% (94.3%).

Multiple genome assemblies were available for *L. acidophilus*, *L. brevis*, *L. delbrueckii*, *L. fermentum*, *L. helveticus*, *L. johnsonii*, *L. paracasei*, *L. plantarum*, *L. reuteri* and *L. rhamnosus*. For these lactobacilli, sequence similarity within the same species ranged from 95.25-100.0% for region V1-V2, with the lowest sequence similarity found between one of the 16S rRNA gene copies in *L. acidophilus* strain 30SC and all copies of the gene in strains NCFM and La-14. For region V3-V4, sequence similarity within the same species ranged from 98.1-100.0%, with the lowest sequence similarity found between one of the 16S rRNA gene copies in *L. plantarum* strain 16 and one of the gene copies in strain p-8. For region V6, sequence similarity between different strains of the same species ranged from 94.3-100.0%, with the lowest sequence similarity found between one of the 16S rRNA gene copies in *L. delbrueckii* strain ATCC11842 and the strains ATCC BAA-365 and 2038.

Between-species sequence similarity

The highest inter-species similarity for the V1-V2 region was between *L. gallinarum* strain ATCC33199 and *L. helveticus* strains H9 and DPC4571 and some copies from strain CNRZ32 whose sequences were identical. The highest inter-species similarity for the V3-V4 region was between *L. gallinarum* strain ATCC33199 and *L.*

crispatus strain ST1 whose sequences were identical. In the V6 region, the highest inter-species similarity was between *L. gasseri* and *L. johnsonii* whose sequences were identical, and between *L. rhamnosus* and *L. paracasei* where all strains shared at least one identical sequence.

Using a cut-off of 97%, the following species could not be separated using the V1-V2 region: *L. gasseri* and *L. johnsonii* (with up to 98.8% inter-species similarity) as well as *L. acidophilus*, *L. crispatus*, *L. gallinarum* and *L. helveticus*. In the latter group, some similarities fell below the 97% cut-off and using a more stringent cut-off of 98% allowed separation of all species within this group, with the exception of *L. gallinarum* and *L. helveticus*.

By comparison, using a cut-off of 97%, the following species could not be fully separated using the V3-V4 region: *L. rhamnosus* and *L. paracasei* (with up to 99.8% inter-species similarity), *L. reuteri* and *L. vaginalis* (with up to 99.8% inter-species similarity), *L. gasseri* and *L. johnsonii* (with up to 98.8% inter-species similarity), as well as *L. acidophilus*, *L. crispatus*, *L. delbruekii*, *L. gallinarum*, *L. helveticus* and *L. jensenii*. In the latter group, some similarities fell below the 97% cut-off, particularly in the case of *L. jensenii*. Using a more stringent cut-off of 98% allows separation of *L. delbruekii* and *L. jensenii* from each other and from the rest of the group.

Finally, in the V6 region, using a cut-off of 97%, the following species could not be fully separated: *L. rhamnosus* and *L. paracasei* (with up to 100% inter-species similarity), *L. gasseri* and *L. johnsonii* (with 100% inter-species similarity), as well as *L. reuteri* and *L. vaginalis* (with up to 98.4% inter-species similarity).

Appendix F: Rarefaction Depth for VMB-HARP

In order to find the rarefaction depth that achieves a good balance between maximising sample retention whilst discarding unreliable sample profiles, two approaches were used. First, for the subset of sample DNA extracts that had been PCR amplified and sequenced twice, the minimum read count of each sample pair was compared with the Bray Curtis similarity between pairs (see Figure F.1). Secondly, rarefaction curves were generated to determine at what sampling depth, the increase in discovered OTUs levelled off (see Figures F.2 and F.3). A read cut-off of 1000 reads was deemed sufficient to ensure adequate sample quality and sequencing depth while also retaining a good number of samples.

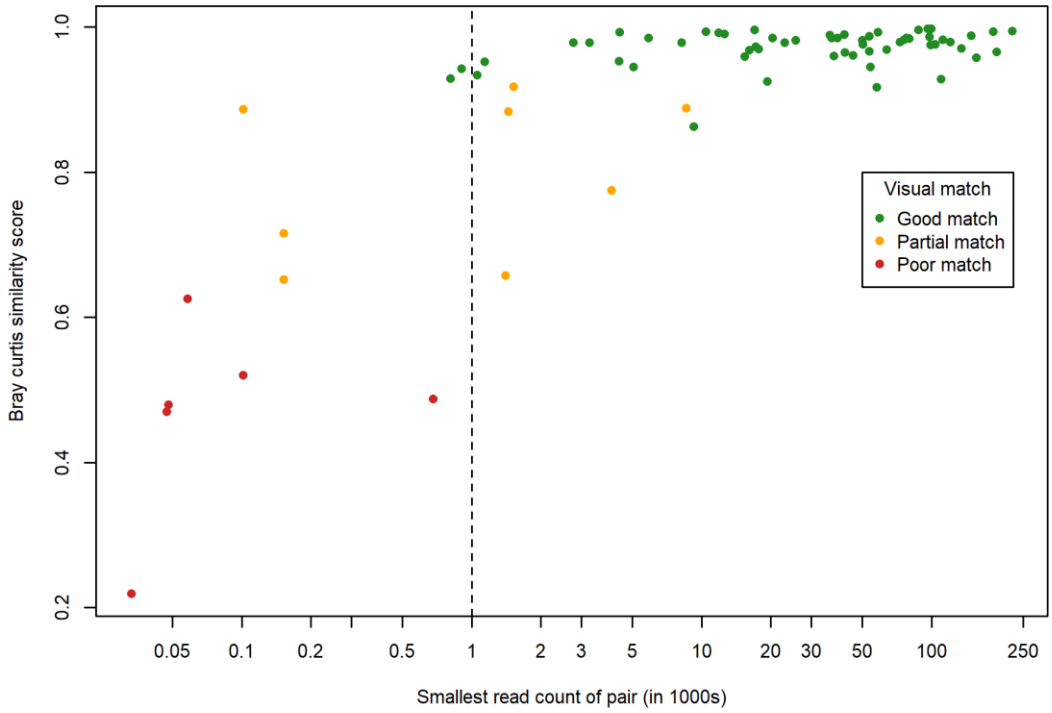


Figure F.1 Scatterplot showing repeatability of all vaginal sample extracts which underwent PCR and subsequent sequencing twice. The lower read count of the pair after removal of contaminant taxa is plotted against the Bray Curtis similarity score between the sample pair (where 0 is no match at all and 1 is a complete match). Data points are coloured by a subjective visual score which was assigned depending on how well the two sample profiles matched on a barchart. A partial match was defined as good concurrence in terms of taxa, but with obvious differences in proportions. The dotted line represents a read count of 1,000. Graph generated in R (version 3.2.2).

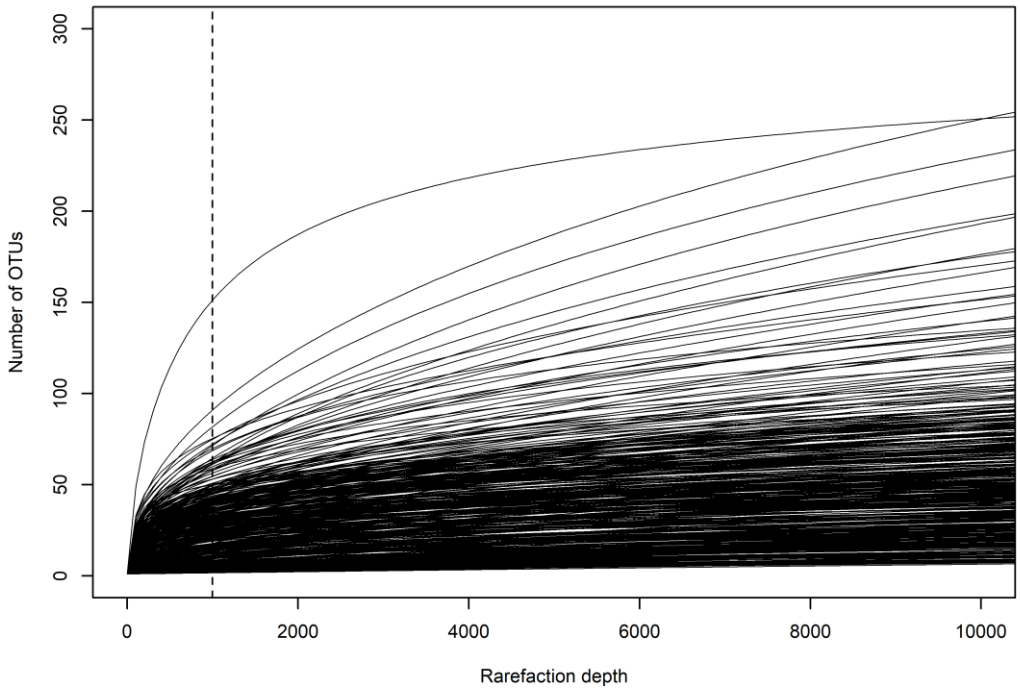


Figure F.2 Rarefaction curve of VMB-HARP samples from visit 1. The dotted line represents a read count of 1,000. Graph generated using the vegan package version 2.3-2 in R (version 3.2.2).

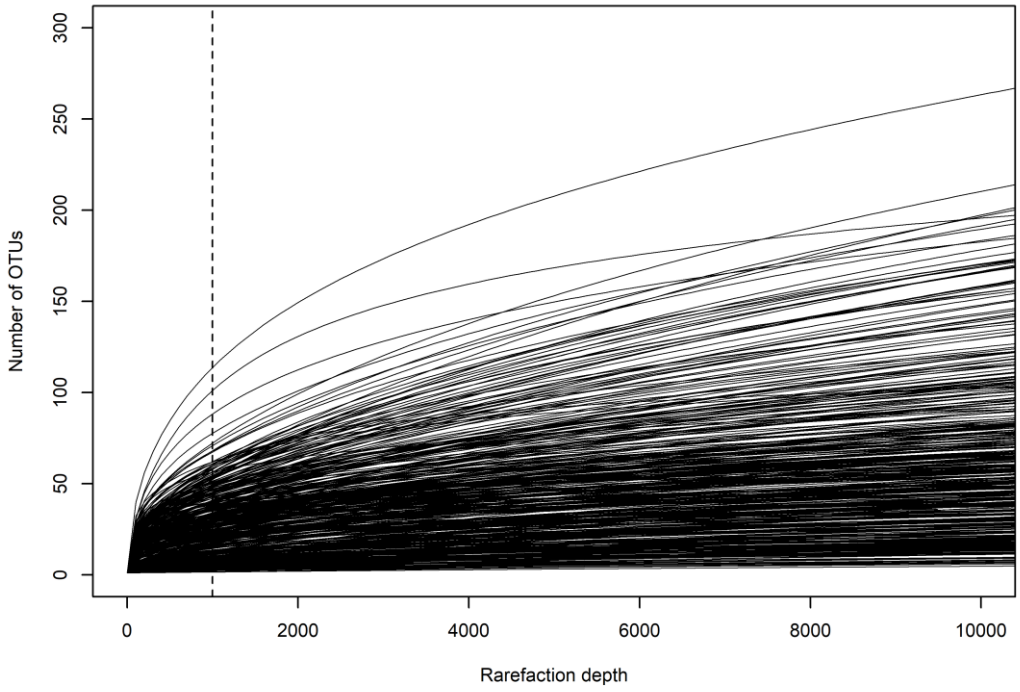


Figure F.3 Rarefaction curve of VMB-HARP samples from visit 5. The dotted line represents a read count of 1,000. Graph generated using the vegan package version 2.3-2 in R (version 3.2.2).

Appendix G: Sequencing Contaminants

PCR and DNA extraction contaminants in the VMB-HARP sequencing data were identified as described in section 4.3.6. Table F.1 lists PCR contaminants in order of decreasing abundance in controls. Table F.2 lists extraction contaminants in order of decreasing abundance in controls. The abundance of each contaminant OTU is given, as a percentage of total reads identified as contaminants within the set of controls.

Table G.1 OTUs identified as PCR contaminants. The abundance of each contaminant OTU is given, as a percentage of total reads identified as contaminants within all negative PCR controls.

PCR Contaminants	
Achromobacter denitrificans/ruhlandii/xylosoxidans	55.37%
Stenotrophomonas maltophilia	24.79%
Delftia acidovorans/lacustris/tsuruhatensis	10.74%
Propionibacterium acnes/avidum	5.79%
Pseudomonas azotoformans/cedrina/fluorescens/gessardii/grimontii/libanensis/marginalis/poae/putida/reactans/syncyanea/synxantha	3.31%

Table G.2 OTUs identified as extraction contaminants. The abundance of each contaminant OTU is given, as a percentage of total reads identified as contaminants within all negative extrraction controls. OTUs that were also identified as PCR contaminants, and may originate from this step are starred.

Extraction contaminants	
Rhodanobacter glycinis/terrae	88.55%
Pseudoalteromonas mariniglutinosa/prydzensis/tetraodonis	4.90%
Vibrio metschnikovii	1.38%
Flavobacterium sp.	0.76%
Pseudomonas mendocina/pseudoalcaligenes	0.38%
Aeromonas caviae/dhakensis/hydrophila/jandaei/media/salmonicida/veronii	0.31%
Pseudarcicella sp.	0.21%
Rhodoluna limnophila	0.20%
Elizabethkingia sp.	0.18%
Saccharibacteria sp.	0.13%
Pseudomonas alcaligenes/alcaliphila/asplenii/chengduensis/indoloxydans/jessenii/mendocina/nitroreducens/oleovorans/pseudoalcaligenes/toyotomiensis/trautweinii	0.13%
Flavobacterium sp.	0.12%
Flavobacterium sp.	0.10%
Stenotrophomonas maltophilia	0.10%
Flavobacterium sp.	0.10%
Chryseobacterium sp.	0.09%
Anoxybacillus flavithermus/kaynarcensis/tunisiense	0.09%
*Stenotrophomonas maltophilia**	0.08%
Shewanella sp.	0.08%
Fluviicola sp.	0.07%
Sporichthyaceae sp.	0.07%

*Propionibacterium acnes/avidum**	0.06%
Streptococcus sp. FF10	0.06%
Chryseobacterium sp.	0.06%
Streptococcus alactolyticus/equinus/gallolyticus/macedonicus/pasteuri/ Pasteurianus	0.05%
Fluviicola sp.	0.05%
Stenotrophomonas sp. 26	0.05%
Stenotrophomonas maltophilia	0.05%
Streptococcus sp.	0.04%
Cytophaga sp.	0.04%
Curvibacter sp.	0.04%
Curvibacter lanceolatus	0.04%
Pseudomonas sp.	0.04%
Cytophaga sp.	0.04%
Chryseobacterium sp.	0.04%
Flavobacterium sp.	0.04%
Elizabethkingia meningoseptica/miricola	0.04%
Limnohabitans curvus/planktonicus	0.04%
*Pseudomonas azotoformans/cedrina/fluorescens/gessardii/grimontii/libanensis/ marginalis/poae/putida/reactans/syncyanea/synxantha**	0.04%
Microbacteriaceae sp.	0.04%
Chryseobacterium sp.	0.04%
Stenotrophomonas maltophilia	0.03%
Paludibacter sp.	0.03%
Sporichthyaceae sp.	0.03%
Cloacibacterium sp.	0.03%
Streptococcus sp.	0.03%
Elizabethkingia sp.	0.03%
Enterococcus cecorum	0.02%
Pseudomonas veronii	0.02%
Polynucleobacter asymbioticus/duraquae	0.02%
Acetobacteroides sp.	0.02%
Saccharibacteria sp.	0.02%
Stenotrophomonas acidaminiphila/humi/maltophilia	0.02%
Legionella sp.	0.02%
Flavobacterium sp.	0.02%
Fluviicola sp.	0.02%
Comamonadaceae sp.	0.02%
Fluviicola sp.	0.02%
Methylobacterium aminovorans/extorquens/organophilum/podarium/populi/ pseudosasae/rhodesianum/suomiense/thiocyanatum/zatmanii	0.02%
Sphingomonas sp.	0.02%
Acinetobacter baumannii/junii	0.02%
Sphingobacteriales sp.	0.02%
Undibacterium sp.	0.02%
Methylobacterium sp.	0.02%
Microbacteriaceae sp.	0.02%
Rhodanobacter sp.	0.02%
Flavobacterium sp.	0.02%
Clostridium sensu stricto 9	0.02%
Holophagaceae sp.	0.02%
Brevundimonas diminuta/naejangsanensis/vancanneytii	0.02%
Paludibacter sp. 70	0.02%
Limnohabitans sp.	0.02%
Sporichthyaceae sp.	0.02%

Flavobacterium sp.	0.02%
Sphingobacteriales NS11-12 marine group	0.02%
Oligoflexales 0319-6G20	0.02%
Sphingobacteriaceae sp.	0.02%
Corynebacterium sp.	0.01%
Methylophilus sp.	0.01%
Acinetobacter gyllenbergii/tjernbergiae	0.01%
Acinetobacter sp.	0.01%
Bacteria sp.	0.01%
Saccharibacteria sp.	0.01%
Peptostreptococcaceae sp.	0.01%
Pedobacter boryungensis	0.01%
Gracilibacteria sp.	0.01%
Spirochaeta 2	0.01%
Collimonas sp.	0.01%
Candidatus Campbellbacteria sp.	0.01%
Sediminibacterium sp.	0.01%
Elusimicrobia Lineage IIb sp.	0.01%
Limnohabitans sp.	0.01%
Neisseriaceae sp.	0.01%
Tepidimonas arfidensis	0.01%
Neisseriaceae sp.	0.01%
Dechloromonas sp.	0.01%
Flavobacteriaceae sp.	0.01%
Chryseobacterium sp.	0.01%
Acinetobacter sp.	0.01%
Chryseobacterium sp.	0.01%
Sphingomonas sp.	0.01%
Brevundimonas sp.	0.01%
Pseudomonas sp.	0.01%
Paludibacter sp.	0.01%
Thermicanus sp.	0.01%
Flectobacillus sp.	0.01%
Candidatus Campbellbacteria sp.	0.01%
Pseudomonas sp.	0.00%
Clostridiaceae 1	0.00%
Pelomonas sp.	0.00%
Alphaproteobacteria sp.	0.00%

Appendix H: Fine Scale Vaginal Microbiome Cluster Description in VMB-HARP Study

Table H.1 Fine scale vaginal microbiome cluster descriptions in VMB-HARP study. Right hand columns represent the number of samples in each cluster at visit 1 and visit 5.

VMB type	Cluster	Description	Visit 1	Visit 5	Total
Lcj	D1	These samples are dominated by *Lactobacillus acidophilus*/*casei*/*crispatus*/*gallinarum* (63-100%).	22	30	52
	D2a	These samples all contain *Lactobacillus acidophilus*/*casei*/*crispatus*/*gallinarum* (44-74%) with a lower relative abundance of *Lactobacillus iners* (25-41%).	6	6	12
	G	These samples are dominated by *Lactobacillus jensenii* (77-98%) together with variable proportions of *Lactobacillus acidophilus*/*casei*/*crispatus*/*gallinarum* (0.3-15%)	0	4	4
Li	Ca	These samples are dominated by *Lactobacillus iners* (75-100%).	105	87	192
	Cc	These samples all contain *Lactobacillus iners* (44-75%) with a lower relative abundance of *Lactobacillus acidophilus*/*casei*/*crispatus*/*gallinarum* (19-46%).	12	10	22
BD	A13	This sample is dominated by *Bifidobacterium breve* cluster 0 (69%) together with *Streptococcus mitis* group (9%), *Prevotella bivia* cluster 0 (5%), an *Escherichia*/*Shigella* sp. (4%) and a *Veillonella* sp (3%).	0	1	1
	A15	This sample is dominated by *Bifidobacterium longum* (85%) together with *Streptococcus anginosus* group (6%) and *Streptococcus pyogenes* group (which includes group B streptococcus; 7%)	0	1	1
L+A	A3	These samples all contain *Lactobacillus crispatus*/*gasseri*/*helveticus*/*johnsonii*/*kefiranofaciens* (42-53%) in combination with *Gardnerella vaginalis* clusters 0 (0.02-23%) and 1 (0.07-43%) and *Peptoniphilus asaccharolyticus*/*grossensis*/*harei* (0.02-2%).	1	4	5
	A4a	These samples all contain *Lactobacillus iners* (27-52%). in combination with *Gardnerella vaginalis* clusters 0 (0.3-23%) and 1 (0.06-12%) and *Atopobium vaginae* cluster 0 (0.02-8%). No other bacterial taxa are consistently present in all samples above 0.01%. However, other common taxa included *Gardnerella vaginalis* cluster 2 (≤39%; present in 33% of samples at >5%), *Megasphaera* cluster 0 (≤25%; present in 38% of samples at >5%), BVAB1 (≤26%; present in 31% of samples at >5%) and *Sneathia amnii*/*sanguinegens* (≤20%; present in 31% of samples at >5%)	22	17	39

	A4b	These samples all contain *Lactobacillus iners* (14-56%). in combination with *Gardnerella vaginalis* cluster 0 (7-48%). No other bacterial taxa are consistently present in all samples above 0.01%. However *Megasphaera* cluster 0 was also common (≤26%; present in 47% of samples at >5%).	19	19	38
	Cb	These samples all contain *Lactobacillus iners* (40-84%), often in combination with *Gardnerella vaginalis* cluster 0 (≤27%; present in 49% of samples at >5%).	60	55	115
	D2b	These samples all contain *Lactobacillus acidophilus*/*casei*/*crispatus*/*gallinarum* (40-61%) with a lower relative abundance of *Gardnerella vaginalis* (37-44%) and a small proportion of *Lactobacillus fermentum*/*gasseri*/*reuteri*/*vaginalis* (0.02-7%).	4	0	4
	F2a	These samples are contain *Gardnerella vaginalis* clusters 0 (0.4-12%), and 1 (35-55%), in addition to variable proportions of *Lactobacillus iners* (0.03-43%) and a small proportion of *Dialister microaerophilus* (0.07-1.4%)	4	1	5
	F2b	These two samples are dominated by *Gardnerella vaginalis* cluster 1 (38 and 44%), and *Lactobacillus jensenii* (28 and 53%). Other bacteria include *Atopobium vaginae* cluster 0 (1.4 and 9%) and *Gardnerella vaginalis* cluster 0 (1.5 and 5%).	1	1	2
BV	A1a	This cluster contains high diversity samples which all contain *Gardnerella vaginalis* cluster 0 (0.03-20%. No other bacterial taxa are present consistently above 0.01%. However, common taxa include *Sneathia amnii*/*sanguinegens* (≤37%; present in 75% of samples at >5%), *Lactobacillus iners* (≤32%; present in 40% of samples at >5%), *Atopobium vaginae* cluster 0 (≤30%; present in 32% of samples at >5%), *Megasphaera* cluster 1 (≤33%; present in 29% of samples at >5%), *Dialister* cluster 0 (≤19%; present in 28% of samples at >5%), *Gardnerella vaginalis* cluster 1 (≤36%; present in 25% of samples at >5%) and *Prevotella* cluster 0 (≤24%; present in 25% of samples at >5%)	52	47	99
	A1b	This cluster contains high diversity samples which all contain BVAB1 (8-42%), *Gardnerella vaginalis* cluster 0 (0.3-16%), *Gardnerella vaginalis* cluster 1 (0.03-4%) and *Dialister microaerophilus* (0.08-1.4%). No other bacterial taxa are present consistently above 0.01%. However, other common taxa included *Lactobacillus iners* (≤34%; present in 52% of samples at >5%), *Megasphaera* cluster 0 (≤21%; present in 48% of samples at >5%), *Prevotella* cluster 0 (≤15%; present in 42% of samples at >5%), *Megasphaera* cluster 1 (≤16%; present in 38% of samples at >5%), *Dialister* cluster 0 (≤29%; present in 32% of samples at >5%) and *Sneathia amnii*/*sanguinegens* (≤24%; present in 32% of samples at >5%)	27	23	50

	A1c	This cluster contains high diversity samples which all contain *Megasphaera* cluster 0 (8-46%), *Gardnerella vaginalis* cluster 0 (0.4-24%), *Gardnerella vaginalis* cluster 1 (0.7-21%), *Atopobium vaginae* cluster 0 (0.5-8%), *Dialister microaerophilus* (0.02-0.6%) and *Veillonellaceae* (0.02-0.5%). No other bacterial taxa are present consistently above 0.01%. However, other common taxa included *Prevotella* cluster 0 (≤32%; present in 63% of samples at >5%), *Sneathia amnii*/*sanguinegens* (≤18%; present in 63% of samples at >5%), *Lactobacillus iners* (≤15%; present in 50% of samples at >5%), *Sneathia sanguinegens* (≤21%; present in 38% of samples at >5%) and BVAB2 (≤14%; present in 25% of samples at >5%).	7	1	8
	A1e	These samples contain *Veillonella montpellierensis* clusters 0 (0.07-30%) and 1 (5-39%), *Lactobacillus iners* (0.03-5%), *Gardnerella vaginalis* cluster 0 (0.3-9%), *Gardnerella vaginalis* cluster 1 (0.3-27%), *Prevotella bivia* (0.04-23%), *Peptostreptococcus anaerobius* (0.18-3%) and *Streptococcus agalactiae/pyogenes* (0.04-7%)	3	3	6
	A2	The main OTU in these samples is BVAB1 (33-68%). Additionally, all samples contain *Dialister* cluster 0 (0.18-8%) and *Dialister microaerophilus* (0.06-0.6%). No other bacterial taxa are consistently present in all samples above 0.01%. However, other common taxa included *Megasphaera* clusters 0 (≤21%; present in 67% of samples at >5%) and 1 (≤23%; present in 33% of samples at >5%) and *Gardnerella vaginalis* cluster 0 (≤36%; present in 47% of samples at >5%).	14	16	30
	A6	These samples all contain *Atopobium vaginae* cluster 0 (27-34%) in combination with *Gardnerella vaginalis* clusters 0 (7-12%), 1 (1.6-8%) and 2 (20-42%).	1	2	3
	A7	These samples consist mainly of *Sneathia amnii/sanguinegens* (26-74%) and *Sneathia sanguinegens* (0.16-22%). Additionally, there are smaller proportions of *Gardnerella vaginalis* cluster 1 (0.13-2%), *Dialister* cluster 0 (0.36-6%), *Dialister microaerophilus* (0.05-1.3%) and *Senegalimassilia* (0.33-5%). No other bacterial taxa are consistently present in all samples above 0.01%. However, other common taxa included *Prevotella* cluster 0 (≤15%; present in 50% of samples at >5%) and *Megasphaera* cluster 1 (≤11%; present in 29% of samples at >5%).	8	6	14
	A8	These samples consist mainly of *Sneathia amnii/sanguinegens* (5-11%) and *Sneathia sanguinegens* (38-57%). Other taxa include *Lactobacillus iners* (0.15-11%), *Gardnerella vaginalis* cluster 0 (1.3-6%) and *Megasphaera* cluster 1 (1.2-12%).	3	0	3
	A9	These samples consist of *Mycoplasma hominis* (38 and 60%), an *Escherichia*/*Shigella* sp. (5 and 13%), *Gardnerella vaginalis* cluster 0 (5%) and *Ureaplasma parvum*/*urealyticum* (3 and 4%).	1	1	2

	A10	The main OTU in these samples is *Prevotella bivia* cluster 0 (37-55%). No other bacterial taxa are consistently present in all samples above 0.01%. Other bacteria include *Anaerococcus* spp., *Atopobium vaginae*, *Bifidobacterium* spp., *Dialister* spp, *Gardnerella vaginalis*, *Megasphaera*, *Mycoplasma hominis*, *Parvimonas* sp., *Peptostreptococcus anaerobius*, *Senegalimassilia* sp., *Sneathia* spp. and *Veillonella montpellierensis*.	4	1	5
	B1a	These samples consist mainly of *Gardnerella vaginalis* clusters 0 (6-57%) and 1 (0.03-21%). No other bacterial taxa are consistently present in all samples above 0.01%. However, other common taxa included *Atopobium vaginae* cluster 0 (≤31%; present in 42% of samples at >5%) Gardnerella vaginalis cluster 1 (≤21%; present in 27% of samples at >5%), *Megasphaera* clusters 0 (≤22%; present in 34% of samples at >5%) and 1 (≤25%; present in 37% of samples at >5%), *Dialister* cluster 0 (≤21%; present in 27% of samples at >5%), *Sneathia amnii*/*sanguinegens* (≤24%; present in 40% of samples at >5%), *Lactobacillus iners* (≤24%; present in 30% of samples at >5%) and BVAB1 (≤25%; present in 27% of samples at >5%).	32	51	83
AD	B1b	These samples consist mainly of *Gardnerella vaginalis* clusters 0 (35-61%) and 1 (11-47%).	6	3	9
	B2	These samples consist mainly of *Gardnerella vaginalis* cluster 0 (42-97%). No other bacterial taxa are consistently present in all samples above 0.01%. However, other common taxa included *Atopobium vaginae* cluster 0 (≤16%; present in 36% of samples at >5%), *Gardnerella vaginalis* cluster 1 (≤10%; present in 32% of samples at >5%) and *Lactobacillus iners* (≤26%; present in 28% of samples at >5%).	15	10	25
	E	These samples all contain *Gardnerella vaginalis* clusters 0 (1.7-3%), 1 (1.0-9%) and 2 (49-84%).	2	1	3
	F1	These samples are dominated by *Gardnerella vaginalis* clusters 0 (0.6-10%), and 1 (60-93%).	6	10	16
	I	These samples are dominated by *Atopobium vaginae* cluster 0 (55-95%), together with variable proportions of *Lactobacillus iners* (0.01-34%). Two samples also contain *Gardnerella vaginalis* clusters 0 (11 and 13%) and 1 (11 and 33%).	0	3	3
PB	A1d	These samples consist mainly of *Lactobacillus iners* (15-47%) and *Streptococcus mitis* group (19-44%). Additionally, there are smaller proportions of *Gardnerella vaginalis* cluster 0 (0.03-1.3%), *Gardnerella vaginalis* cluster 1 (0.05-0.13%), *Sneathia sanguinegens* (0.02-0.05%), *Granulicatella elegans* (0.06-1.3%) and a *Staphylococcus* sp. (0.02-0.09%).	3	0	3
	A5	This sample contained *Streptococcus anginosus*/*milleri* (59%) and *Lactobacillus crispatus*/*gasseri*/*helveticus*/*johnsonii*/*kefiranofaciens* (38%).	0	1	1

	A14	These samples contain an *Escherichia*/*Shigella* sp. (43-74%), *Streptococcus anginosus* group (7-33%), *Enterococcus durans*/*faecalis*/*faecium* (1.2-7%), *Gardnerella vaginalis* cluster 1 (0.1-14%) and *Streptococcus* mitis group (0.4-13%).	1	1	2
	H1	These samples all contain *Streptococcus agalactiae*/*pyogenes* (29-54%). No other bacterial taxa are consistently present in all samples above 0.01%. However, other common taxa included *Prevotella bivia* (≤31%; present in 60% of samples at >5%) and *Gardnerella vaginalis* cluster 1 (≤34%; present in 60% of samples at >5%).	1	4	5
	H2	These samples are dominated by *Streptococcus agalactiae*/*pyogenes* (53-98%), together with variable proportions of *Lactobacillus iners* (0.01-34%).	3	4	7
	J	This sample is dominated by *Streptococcus dysgalactiae*/*pyogenes* (98%).	0	1	1
N/A	A11	This sample is dominated by *Scardovia wiggsiae* (79%) together with *Streptococcus anginosus* group (12%).	0	1	1